Statistical methods for EEG data

José Cortiñas Abrahantes
EFSA European Food Safety Authority

Helena Geys
Hasselt University

Geert Molenberghs
Universiteit Hasselt and University of Leuven

Wilhelmus H Drinkenburg
Janssen Pharmaceutica

Statistical Methods for EEG Data

José Cortiñas Abrahantes, Jan Serroyen, Helena Geys, Geert Molenberghs

Limburgs Universitair Centrum, Center for Statististics, Universitaire Campus, B3590
Diepenbeek, Belgium

Wilhelmus H.I.M. Drinkenburg

Johnson & Johnson Pharmaceutical Research and Development, Turnhoutseweg 30,
B2340 Beerse, Belgium

1 Introduction

We will first introduce the area of psychotropic drugs, then discuss pharmaco-electroencephalogram studies, whereafter we will indicate how they pose a problem from a methodological point of view and how initial answers have been formulated. In subsequent sections, particular statistical methodology will be introduced and exemplified using two case studies.

1.1 Psychotropic Drugs

For thousands of years, humans in all known societies have used psychotropic drugs (substances that act on the central nervous system (CNS) and affect mood, thinking, and behavior). Psychotropic drugs, whatever the substances used, rank second to tenth among the most consumed medicine in Western nations (Zarifian 1996). Their goal is to modify (to increase or to reduce) the cerebral transmissions carried out by neurotransmitters (such as the dopamine) whose dysfunction could be at the origin of mental disorders (Costentin 1993).

Roughly speaking, all psychotropics may be classified as CNS depressants (e.g., alcohol, opium, Valium), CNS stimulants (e.g., coffee, cocaine, Ritalin), hallucinogens (e.g., LSD), etc. Psychotropics either tranquillize or sedate, awake or stimulate, or impair perception. Occasionally, shades of these effects are produced by a single drug.

Over the last 45 years, with progress in synthesizing chemicals and with the rapid acceptance of medical approaches to understand and treat human distress, many psychotropics (except, since the 1960s, hallucinogens) have been prescribed to treat emotional and behavioral problems. These drugs are often called "psychiatric drugs," although primary care physicians write the bulk of prescriptions.

In spite of recurrent controversies (Zarifian 1988, Kapsenbelis 1994), these substances may be divided into 5 major classes (Deniker 1982, Oughourlian 1984, Cohen and Cailloux-

Cohen 1995), each typically named according to its main indication in psychiatry. However, these names mean little since most drugs are actually used for most indications: stimulants may be prescribed to calm children, antidepressants to relieve anxiety; anticonvulsants to control mania, etc. Classifying drugs on the basis of known chemical structure or action might be more accurate but would create numerous categories. Given the absence of pathophysiological findings underlying psychiatric diagnoses, current classification of psychiatric drugs remains provisional.

1.2 Pharmaco-electroencephalogram Studies

Pharmaco-electroencephalogram (EEG) studies aim at characterizing psychotropic drug effects usually on the basis of spectral EEG analysis. This way EEG-defined sleep-waking behavior can be explored and, in conjunction with electromyogram (EMG) and movement monitoring, clearly defined states of vigilance can be separated out, resulting, e.g., in a hypnogram.

Typically, six sleep-wake stages are distinguished: (1) *active wake*, characterized by movement, theta activity and high EMG, (2) *quiet wake*, without movement, (3) *light sleep*, characterized by EEG spindles, (4) *deep sleep*, with slow waves and prominent delta activity, (5) *intermediate sleep*, with spindles against a background of theta activity and low EMG, and (6) *paradoxical sleep*, with theta activity and low EMG. We will use the following abbreviations: active wake (AW), quiet wake (QW), light sleep (Sws1), deep sleep (Sws2), intermediate sleep (IS) and paradoxical sleep (PS).

Experiments typically allocate n rats over different treatment groups, generally several doses of the same drug and a placebo group are the type of experiments encountered in this area. The brain signals of the rats are monitored during a number of hours, which generally are divided into a light period and a period of not light. The administration of the drug is usually done at the beginning of the light period and after each experiment a period of washout is considered in order to use the same rat again in another experiment. The effects of the drugs on sleep-waking behavior are assessed using several parameters from the hypnogram. In the application of Section 8, we will focus on the time spent in each sleeping stage as a summary measure, and we refer to the sections on longitudinal data for the situation where time is taken into account as well.

1.3 A statistical Challenge

From a statistical point of view, analyzing EEG data poses rather a though challenge. This is due to several reasons. First, there is the high-dimensionality of raw EEG data. Even after the usual initial reduction of dimensionality involving a spectral analysis, in which the power spectrum is divided into several frequency bands (*delta, theta*, etc.), there is a multitude of variables to be analyzed.

Secondly, the fact that a pharmaco-EEG study usually consists of subjects that are measured repeatedly over time indicating that we are dealing with *longitudinal data*. Longitudinal data require special statistical methods because the set of observations on one subject tends to be intercorrelated. This correlation must be taken into account to draw valid scientific inferences (Diggle, Liang, and Zeger 1994). The *mixed-effects model* is a flexible and widely used approach when modelling this type of data. The theoretical background of mixed-effects models will de discussed later on in Section 4.

Thirdly, there does not exist a reasonably well accepted functional form for the evolution of the EEG activity over time. This immediately becomes clear when looking at e.g. the longitudinal profiles of the time spent in a certain sleep-wake stage. These longitudinal profiles are usually highly irregular and the variability both between and within subjects is relatively high. Often the variance is also not constant over time. Fitting a simple linear model for this kind of data will clearly not be satisfactory. Finding a suitable statistical model that captures the trends of EEG data over time is therefore not a trivial task.

Finally, given these complexities, discriminating between various components in terms of their action is not an easy task. While conventional discriminant analysis methods can be used, a fully satisfactory answer requires more advanced methods to be used and, arguably, further research.

1.4 Solutions from the Field

In order to tackle these statistical challenges, people from the field of pharmaco-EEG have tried to simplify the EEG data even further to be able to use rather basic statistical techniques. To start with, all response variables (e.g., sleep-wake stages) are analyzed separately using univariate statistical methods. Applying multivariate methods would increase the complexity of the analysis considerably.

The most common applied technique is an analysis by time point and correcting for the number of significance tests post-hoc (e.g., Bonferroni correction) in order to deal with the so-called *multiple testing* problem.

Another approach could be to summarize the whole longitudinal sequence of observations into one summary statistic (e.g., average, sum, area under the curve, etc.) per subject and then analyze this summary statistic, e.g., by using ANOVA or a nonparametric test like the Wilcoxon.

However, it is very important to realize that EEG data can be analyzed using a battery of methods which exist for data arising from life science applications, perhaps tailored to the specific needs of the problem at hand. Section 2 provides a broad overview of commonly used methods in biometric applications. The following four sections are devoted to methodology particularly relevant in the area of EEG data: analysis of variance is discussed in Section 3, while random-effects models are highlighted in Section 4; general discrimination and classification methodology is the topic of Section 5, and the specific subfield of classification and regression trees is dealt with in Section 6. An application of the former two sets of methods is presented in Section 7, while the latter two are exemplified in Section 8.

2 Statistical Methods in Biometry

Choosing a statistical approach is a very common task in everyday statistical practice. When choosing a method for analysis, it is important to reflect on whether the methodology is sound from a theoretical point of view and whether it is adequate in terms of the scientific research question of interest. A method chosen should therefore reflect the design, type of outcome, type of covariates, etc. A useful distinction is made between methods for univariate (single) outcomes and methodology for correlated sets of response variables. General texts are Shoukri and Pause (1999), Dunn and Everitt (1991), Pagano and Gauvreau (1992), and Rosner (1994). When the focus is on epidemiological methods, useful references are Breslow and Day (1990) and Rothman and Greenberg (1996).

The simplest statistical analysis is concerned with a single outcome variable, recorded for a sample of a homogeneous population. Standard procedures include the computation of means or medians (location parameters) and standard errors or interquartile ranges (dispersion parameters). For example, the height of a number of human subjects might be recorded. A first level of complexity arises when a variable is recorded for a sample out of two subgroups (subpopulations) of a larger population (treated and untreated patients, two species, boys and girls): the two-sample problems. A question of interest is whether the means are different in the two populations. The outcome variable might still be height, but we would have an explanatory variable: treatment allocation, or sex. For example, the height of boys can be compared to the height of girls. The outcome variable is often called dependent variable. The predictor is often called covariate or independent variable. The statistical tools for this data setting include analysis of variance (ANOVA), t test, Wilcoxon test.

In the previous situation, the dependent variable had only two levels: a binary or dichotomous variable. This is the simplest case. Alternatively, the predictor itself could be a variable with several levels (e.g., dose administered in a clinical trial; one of several species of a plant; race, etc.). In addition, it could potentially have an infinite number of levels, just as is the case with the height response variable. For example, a baseline height at 7 years of age can be compared to the height at 10 years. This leads to a family of models frequently referred to as regression models. When the dependent variable is continuous (height) one often uses linear regression. The independent variable can be continuous, binary, categorical, or discrete. The choice of the statistical analysis method is driven by the outcome or dependent variables, rather than by the predictor variables.

Should the dependent variable be binary (diseased/non diseased; death/alive, etc.), then one would choose logistic regression rather than linear regression. Several alternatives to logistic regression exist, such as probit regression, where an underlying latent normal variable is considered to give rise to the observed binary outcome, after dichotomization, rather than a continuous logistic variable.

Of course, one does not need to be restricted to a single predictor variable. For instance, both treatment allocation and sex of the human subject might be of interest. In such cases, most of the well-known methods easily extend. One-way ANOVA extends to two-way or even multi-way ANOVA. Simple, or single, linear regression extends to so-called multiple regression. Most other techniques, such as logistic regression, are easily extended to encompass multiple covariates. It has to be noted that, while simple in theory, methods for multiple covariates require great care since particular issues are raised that do not occur otherwise. Indeed, such issues as collinearity arise only for multiple covariate models. Often, not all predictors are on equal footing. For example, the relation between an exposure and a disease is of interest, while another variable is merely a confounder. This issue needs careful consideration in all non-randomized settings, such as epidemiological or otherwise observational studies. Thus, model building and interpretation of (regression) coefficients require both expertise as well as subject matter knowledge.

Particular care is needed in cases where the outcome variable is a time to a certain event. In a life sciences context, this is often the time from the beginning of a study, birth, or start of randomization, until a certain medical event occurs, such as death, relapse of onset of disease, complete cure, pregnancy, etc. This methodological area is often referred to as survival analysis or lifetime data analysis. There are two main reasons why standard (linear) regression is seldom appropriate. First, survival times tend to show skew rather than symmetric distributions, unlike in the normal distribution. Second, and more important, is the potential occurrence of censoring, i.e., the follow up time for a subject is not sufficiently long in order to observe the actual survival time. In such a case, it is clear that the actual survival times exceeds the end of follow-up, i.e., the survival time is larger than the censoring time. This means that partial, or coarse, information is present. Nevertheless, such information needs to be included into the analysis. A lot of research has been devoted to develop parametric and non-parametric methods for the analysis of survival times in the presence of censoring.

Another important set of situations, different from the univariate settings considered thus far, occurs when several dependent variables are recorded simultaneously. This concept harbors a large, and ever growing portion of statistical methodology, of use in health sciences and elsewhere. The most classical setting is multivariate analysis, where different outcomes are measured on the same subject. Alternatively, the same measurement can be taken on the same unit or on correlated units. Examples include longitudinal studies, where the same measurement is repeatedly made over time, spatial statistics, where the connection of a response to its geographical location is of interest, clustered data (e.g., in animal litters or in family studies), hierarchical survey data, such as arising from multistage or cluster sampling, etc. A particular area of such dependent, or hierarchical, data is given by meta-analysis in clinical trials, where information is pooled from several clinical trials. Apart from general methodological considerations, one is then confronted with pragmatic issues such as which study to include, etc.

Section 2.1 is devoted to classical univariate modeling, including linear regression, generalized linear models, and logistic regression. Section 2.2 offers a perspective on hierarchical data, encompassing multivariate methods as well as repeated measures and multilevel modeling. Finally, some comments about survival analysis are made, and its connection to the other modeling frameworks is highlighted.

2.1 Linear Regression, Generalized Linear Models, Exponential Family, and Logistic Regression

2.1.1 Gaussian Outcomes

The analysis of continuously distributed responses (Neter *et al.* 1996), especially when they are normally distributed, has received a lot of attention. Next to the t test, analysis of variance and linear regression have received a lot of attention. The general linear regression model is customarily written as

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i,$$

where Y_i is a response variable for subject $i = 1, \ldots, N$ in a study, X_{ij} is the value for the jth predictor variable and ε_i is an error term. There are some important special cases. For example, when $p = 1$ then there are no covariates and the one-sample problems results. When $p = 2$, so-called simple or single regression is obtained, where the outcome variable is regressed on a single covariate. When all of the covariates are dummy variables (0 or 1 depending on whether a certain characteristic is absent or present within a subject), possibly resulting from a multi-categorical covariate, then analysis of variance is obtained. Analysis of variance and regression are often treated as different entities in introductory texts. This makes sense because on the one hand linear regression generalizes ANOVA, while on the other hand a larger number of results and tools is available for the ANOVA setting than for the more general regression setting. In a sense, ANOVA refers to categorical covariates, whereas regression focuses on continuous covariates, or a combination of continuous and categorical covariates.

Regarding the error term, two views can be taken. First, one can restrict attention to specification of its moments only. Most commonly, one assumes a zero mean and a constant variance, σ^2 say. This results in the so-called ordinary least squares (OLS) approach to linear regression. Alternatively, the error term, and subsequently the response variable itself, can be considered normally distributed. In the first case, sampling-based or frequentist inference results, in the second case, full maximum likelihood follows. Both approaches yield the same parameter estimates and almost the same estimates of precision, given that they are asymptotically equivalent. The OLS approach is valid under weaker assumptions, thus even when the errors are not normally distributed, but if one is comfortable with the normal distribution, use can be made of fully parametric inference. This is one of many instances in statistical modelling: if one is prepared to make assumptions, more results become available, but the risk of incorrect assumptions is always present. This is why careful assessment of assumptions is important.

Throughout statistical analysis, and not only in linear regression, the normal distribution is omnipresent. Let us reflect on this phenomenon. For a simple random sample with just one outcome variable, the mean and the standard deviation, and/or the standard error of the mean, are easily computed. This is independent of the true distribution of the data. However, for some distributions, a mean and standard deviation will be less meaningful. This includes, for example, bimodal distributions. Even though they may have a mean, primary scientific interest may lie in identification of the two modes and other characteristics thereof. Another example is provided the Cauchy distribution, which does not have finite mean and variance.

The normal distribution has easy interpretations for the mean and standard deviation of samples drawn from it. The usual definitions of mean and standard deviation are the least squares estimators (and maximum likelihood estimators) of the population quantities. Very importantly, under regularity conditions, the sample mean converges to the location parameter of the true distribution, even if it is not normal. This is based on the so-called law of large numbers (central limit theory), which means, roughly speaking, that distributions of estimators from large samples show a normal spread, even when the samples themselves are drawn from non-normal distributions. In addition, the researcher disposes of an alternative set of tools consisting of transformation methods. This allows to transform responses or residuals that are non-normal, to (more) normally distributed ones. For all of these reasons, the normal distribution is a convenient working paradigm in a number of statistical areas.

2.1.2 Non-Gaussian Outcomes

For the analysis of categorical response variables (Agresti 1990, McCullagh and Nelder 1990), different families of approaches exist. First, when there is one or a few categorical responses, possibly with in addition a set of categorical covariates, so-called contingency table analysis is a commonly used tool. A key model in such a context is the loglinear model. Second, when there is one binary (or, to some extent, categorical) outcome, in the presence of one or more, perhaps continuous, covariates, then appropriate regression tools can be used. In the regression framework, one of the most commonly used tools is logistic regression. There are at two obvious reasons for this. First, it is considered a natural extension of linear regression. Second, especially in a biometrical context, the interpretation of its parameters in terms of odds ratios is considered convenient as it is related to the so-called relative risk. When the latter is less of a concern, such as in econometric applications, one frequently encounters probit regression. The split between these two strands of research is reminiscent of the split between ANOVA and regression methods for continuous outcomes. At the same time, just as these two strands come together in a generic perspective on linear regression, thereby encompassing ANOVA, the theory of generalized linear models encompasses both contingency table analysis (through the loglinear model) as well as logistic regression. Moreover, the framework of generalized linear models is very broad, in the sense that it contains linear regression as a special case, and apart from logistic regression for binary data, it can easily deal with multi-categorical outcomes, whether nominal (unordered) or ordinal (ordered), with counts, and, to some extent, with survival analysis.

Consider a response variable Y_i, measured on subjects $i = 1, \ldots, N$, together with covariates $\boldsymbol{x}_i$. A generalized linear model minimally specifies the mean $E(Y_i) = \mu_i$ and links it to a linear predictor in the covariates $\eta(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}$, where $\eta(.)$ is the so-called link function. Further, the variance of Y_i is then linked to the mean model by means of the mean-variance link

$$\mathsf{Var}(Y_i) = \phi v(\mu_i),$$

where $v(.)$ is a known variance function and ϕ is a scale or overdispersion parameter. Such a specification is sufficient to implement moment-based estimation methods, such as iteratively reweighted least squares or quasi likelihood. In case full likelihood is envisaged, the above framework can be seen to be derived from the general exponential family definition

$$f(y|\theta_i,\phi) = \exp\left\{\phi^{-1}[y\theta_i - \psi(\theta_i)] + c(y,\phi)\right\} \tag{1}$$

with θ_i the natural parameter and $\psi(.)$ a function satisfying $\mu_i = \psi'(\theta_i)$ and $v(\mu_i) = \psi^{''}(\theta_i)$. Hence, the previous results are recovered but extended. From (1) it immediately follows that the corresponding log-likelihood is linear in the statistics θ_i, simplifying the form of the score equations,

$$S(\boldsymbol{\beta}) \;=\; \sum_i \frac{\partial\mu_i}{\partial\boldsymbol{\beta}}\, v_i^{-1}\, (y_i-\mu_i) \;=\; 0,$$

log-likelihood maximization and corresponding statistical inference.

For example, in the case of a binary outcome Y_i, the model can be written as

$$f(y_i|\theta_i,\phi) = \mu_i^{y_i}(1-\mu_i)^{1-y_i} = \exp\left\{y_i \ln\left(\frac{\mu_i}{1-\mu_i}\right) + \ln(1-\mu_i)\right\}$$

and hence the Bernoulli model and, by extension, logistic regression, fits within this framework. In particular,

$$\theta_i = \mathsf{logit}(\mu_i) = \mu_i/(1-\mu_i) = \mathsf{logit}[P(Y_i=1|\boldsymbol{x}_i)], \tag{2}$$

$\mu = e^\theta/(1+e^\theta)$ and $v(\mu) = \mu(1-\mu)$.

In case one opts for a probit link, the logit in (2) is replaced by the inverse of the standard normal distribution Φ^{-1}, i.e., the probit function. This model cannot be put within the exponential family context. Hence, the choice for logistic regression is often based on the mathematical convenience entailed by the exponential family framework. Now, it has been shown repeatedly that the logit and probit link functions behave very similarly, in the sense that for probabilities other than extreme ones (say, outside of the interval $[0.2; 0.8]$) both forms of binary regression provide approximately the same parameter estimates, up to a scaling factor equal to $\pi/\sqrt{3}$, the ratio of the standard deviations of a logistic and a standard normal variable.

The beauty and elegance of the exponential family framework should not disguise that there are fundamental differences with linear regression. First, the normal densities, explicitly or implicitly underlying linear regression, exhibit a separation between mean and variance; this is radically different in most commonly used generalized linear models. Second, the link function introduces a form of non-linearity that is absent in linear regression. This has important consequences in terms of model selection. For example, omitting a covariate in linear regression implies, under certain circumstances, merely an increase in residual variability. In contrast, in GLM, since covariates enter the linear predictor, and the variability is given by the mean-variance link, a non-linear relationship enters into the picture, implying that, when two hierarchically ordered models are considered, at most one of them can be correctly specified.

In spite of these remarks, logistic regression has found its way into everyday statistical practice. Perhaps due to this familiarity, the model has been extended to a number of different settings, including ordinal data, multivariate binary data, and longitudinal data. But, even more so in the extensions, one needs to be aware of fundamental differences between the Gaussian and non-Gaussian settings. This will be exemplified using ordinal data, in the following section.

2.1.3 Regression Models for Ordinal Data

Regression models for binary data, such as described in the previous section, have been extended to nominal and ordinal categorical outcomes. Let us concentrate on ordinal outcomes. Assume that the binary variable $Y_i \in \{0, 1\}$ is replaced by an ordinal one taking values $Y_i \in \{1, 2, \ldots, K\}$. Consider the case of a single covariate x_i. A predictor, linear in the covariate, would take the following form in the binary case:

$$\text{logit}[P(Y_i = 1|x_i)] = \alpha + \beta x_i. \tag{3}$$

A commonly used extension of logistic regression to this case is so-called *proportional odds* logistic regression:

$$\text{logit}[P(Y_i \leq k|x_i)] = \alpha_k + \beta x_i, \qquad k = 1, \ldots, K-1. \tag{4}$$

In (4), the probability of observing a lower response *versus* a higher one is modeled. The term *proportional odds* derives from the fact that the odds for a unit increase in x_i are equal to $\exp\beta$, irrespective of the cutoff, given it nice interpretational properties and elegance *provided the model is correctly specified*. The latter is important and fundamentally different from logistic regression. To see this, consider a logistic regression as in (3) with x_i binary and taking values 0 or 1. For each of the two levels of x_i, there is then one parameter, the probability of success given x_i. Since (3) contains two free parameters, the model is saturated and, in this case, logistic regression is merely a convenient way to model the two probabilities and the difference between them, thereby assuring that for all values of α and β valid (i.e., within the unit interval) probabilities are obtained.

In case of (4), there are $2K - 2$ free probabilities for each of the two levels of x_i, implying that the K free parameters impose model constraints. An obvious extension would be to allow for category dependent effects β_k $(k = 1, \ldots, K-1)$. This model is saturated and can be used as a starting point for model simplification, in this simple contingency table setting.

With continuous covariates, the situation is different. Assuming x_i is continuous, and the fit of model (4) is inadequate (assessed, for example, using a score test, as is routinely done in the SAS procedure GENMOD), one could, in principle, let the covariate effects be category dependent. However, the consequence is that there always exist regions in the covariate space, for any combination of the parameters, where non-valid probabilities would be obtained. Indeed, it is easy to see that the conditions for valid probabilities

$$\alpha_k + \beta_k x_i \leq \alpha_{k+1} + \beta_{k+1} x_i, \qquad k = 1, \ldots, K-1,$$

impose $K-1$ linear inequality constraints. Depending on the signs of $\beta_{k+1} - \beta_k$, the resulting allowable space can be a finite or infinite interval. The only way in which to remove the constraints is by setting the β_k parameters equal, i.e., proportional odds regression.

In case the resulting allowable interval for x_i, for a given set of parameters, corresponds to a scientifically plausible range, the model could still be used. Thus, in general, it is important to realize that there ought to be a careful discussion, when using ordinal data logistic regression, considering the pros and cons in terms of plausibility, flexibility, and constraints.

Of course, (4) is not the only ordinal logistic regression type model. Alternatively, one can consider the multigroup logistic model, where each category is referred to the baseline category. Such a model is mathematically more convenient since it avoids parameter space violations and fits within the exponential family framework, but it does not exploit the ordinal nature of the data. The latter may lead to less parsimonious models and, more importantly, to difficulties in extracting relevant conclusions from the data.

Another approach is to consider *continuation-ratio models*:

$$\text{logit}[P(Y_i > k | Y_i \geq k, x_i)] = \alpha_k + \beta_k x_i, \qquad k = 1, \ldots, K-1. \qquad (5)$$

This model has been given some attention in the literature. Such a model might be convenient and useful for subjects that gradually go through a number of states, where no return is possible (e.g., cancer stages). Fitting the model is easy since (5) consists of $K-1$ separate logistic regressions; only a straightforward expansion of the data is necessary to prepare them for standard calls to logistic regression software.

Nevertheless, while this model might be a convenient option for *directionally* ordered categorical data, it is *not* so when the direction of the ordering is immaterial. This is the case, for example, when a 5-point scale, ranging from 'very bad' to 'very good' can just as well be reversed: 'very good' to 'very bad'. Reversing the coding in such a case merely changes the signs of the parameters involved in the case of proportional odds logistic regression, but it fundamentally changes the model in the continuation-ratio case. Precisely, not only is there no simple transformation between the parameters, significance may change as well and the likelihood at maximum can be different. This is one of the most dramatic instances, in the case of univariate logistic regression for ordinal data, where consideration of a particular model is not just open to criticism, but actually totally meaningless in a number of cases.

2.2 Hierarchical Data

In applied sciences, one is often confronted with the collection of *correlated data*. This generic term embraces a multitude of data structures, such as multivariate observations, clustered data, repeated measurements, longitudinal data, and spatially correlated data (Aerts *et al.* 2002, Verbeke and Molenberghs 2000, Diggle *et al.* 2002, Fahrmeir and Tutz 2001).

Among the hierarchical settings, multivariate data have received most attention in the statistical literature. Techniques devised for this situation include multivariate regression and multivariate analysis of variance, which have been implemented in standard statistical software.

As an example of a simple multivariate study, assume that a subject's systolic and diastolic blood pressure are measured simultaneously. This is different from a *clustered setting* where, for example, for a number of families, diastolic blood pressure is measured for all of their members. A design where, for each subject, diastolic blood pressure is recorded under several experimental conditions is often termed a *repeated measures* study. In the case that diastolic blood pressure is measured repeatedly over time for each subject, we are dealing with *longitudinal data*. Although one could view all of these data structures as special cases of multivariate designs, there are many fundamental differences, thoroughly affecting the mode of analysis. First, certain multivariate techniques, such as principal components, are hardly useful for the other designs. Second, in a truly multivariate set of outcomes, the variance-covariance structure is usually unstructured, in contrast to, for example, longitudinal data. Therefore, the methodology of the general linear model for multivariate data is too restrictive to perform satisfactory data analyses of these more complex data. In contrast, the *general linear mixed model*, is much more flexible.

2.2.1 Multivariate Analysis

Multivariate analysis refers to a set of techniques which allows the presence of more than one outcome variable (Johnson and Wichern 1992, Krzanowski 1988). For example, height and weight might be recorded simultaneously for a group of boys and girls. Arguably, sex will influence height as well as weight. At the same time, height and weight are likely to be correlated or associated. Note that association refers to the concept of dependence between two or more variables. In contrast, correlation refers to a family of measures that can be computed to capture association (Pearson correlation, Spearman correlation). Especially for categorical data, a million measures of association have been proposed as alternatives to the correlation (including the odds ratio, concordance, Kendall's τ, the κ coefficient,. . .).

A different example is provided by the classical Iris Data Set, where four variables (petal length, petal width, sepal length, sepal width) are recorded for 150 irises, subdivided in three equal subsamples of 50 irises, belonging to each of three species: setosa, versicolor, and virginica. Thus, there are four continuous outcomes and one independent variable (species) which is categorical with three levels.

In general, one might have a set of dependent variables, some of which are continuous, discrete, categorical, or binary, and, at the same time a set of independent variables, some of which are continuous, discrete, categorical, and binary. The most general setting is very hard to study. During the last century, a multitude of sub-problems of the general problem have been studied.

It is important to appreciate the relative positions of multivariate analysis on the one hand, and longitudinal, spatially correlated, clustered, or otherwise hierarchical settings. Since correlated responses are recorded, all of these settings differ from univariate analysis. However, in a multivariate study, several variables for simple, as opposed to compound, subjects, such as individuals, are recorded. In contrast, repeated measures, longitudinal data, etc. usually involve compound units (families, litters) or elaborate sampling mechanisms (extending over time: longitudinal; or extending over space: spatial). As a consequence of this difference, models for a multivariate setting will differ from models for the other settings.

In analogy to the univariate setting, the multivariate normal distribution is heavily used in multivariate statistics. This has led to multivariate regression, multivariate analysis of variance, principal components analysis, factor analysis, canonical correlation analysis, discriminant analysis, the biplot, multidimensional scaling, etc.

For the same reasons as in the univariate setting, these techniques can be used also when samples are non-normal, for two reasons. First, merely for describing, summarizing, reducing data (such as a sample mean), no (full) distributional assumptions have to be made. Second, for samples large enough, asymptotic theory can be invoked to prove (normal) properties of the estimators' sampling distributions, even when the data themselves or the residuals thereof are non-normally distributed. Of course, whether a sample is "large" depends on the discrepancy between the true distribution and the normal distribution. Clearly, for binary data the discrepancy might be too large. Therefore, more recent work focuses on non-normal data, particularly with emphasis on multivariate generalized linear models, rather than just multivariate general linear models.

With continuous or quasi-continuous (non-continuous outcomes such as ordinal outcomes with a sufficiently large number of categories) responses, one might still consider a multitude of questions. Consider the Iris data set and let us, for the sake of illustration, focus on three types of question in particular. (1) Are there differences between the species? (2) Given a species, do we need all four variables to have an idea about the characteristics? (3) What is the relation between the sepal variables on the one hand, and the petal variables on the other? Formulated differently, is the sepal shape and/or size related to the petal shape and/or size?

If the first question were asked in a univariate sample, a difference in the mean parameter is usually the subject of investigation. One can test for equality of the mean between the different species. In the current multivariate setting, differences can be spread out over 4 outcomes. Is only one variable responsible for a difference? Is the difference spread out over all four? Suppose we were to classify a new observation. If only one outcome is responsible for the difference, one could confine attention to the value of that one to classify the new observation. However, if it is spread out over all four it is less clear how to do it. Might it be possible to define a new variable, based on the other four, making the job easier? This will be the subject of discriminant analysis.

If the second question were asked in a univariate setting, the information contained in the data for a species is summarized by 1 mean and 1 variance. Here, we have 4 means, 4 variances, and 6 correlations! (Think of 10 outcomes...). Thus, in order to gain insight, a data reduction is needed. This has led to principal components.

To answer the third question, the four variables have been divided into two groups of two variables. In other words, we are interested in the correlation between two *sets* of variables, instead of simply between two variables. Not only are there several correlations between the sets (four in this case), but also the correlation between the variables within each set (one correlation within each set) is present. Rather than having a single correlation, we have six numbers. A satisfactory answer to this question is the subject of canonical correlation analysis. This method generalizes both Pearson's correlation as well as multiple correlation and R^2, known from regression.

2.2.2 Longitudinal and Other Hierarchical Data

Recall that the broad family of correlated data includes longitudinal data, spatially correlated data, clustered data, survey data, etc. Among the clustered data settings, longitudinal data perhaps require the most elaborate modeling of the random variability. Generally, three components of variability can be considered. The first one groups traditional random effects (as in a random-effects ANOVA model) and random coefficients. It stems from interindividual variability (i.e., heterogeneity between individual profiles). The second component, serial association, is present when residuals close to each other in time are more similar than residuals further apart. This notion is well known in the time-series literature. Finally, in addition to the other two components, there is potentially also measurement error. This results from the fact that, for delicate measurements (e.g., laboratory assays), even immediate replication will not be able to avoid a certain level of variation. In longitudinal data, these three components of variability can be distinguished by virtue of both *replication* as well as a clear *distance* concept (time), one of which is lacking in classical spatial and time-series analysis and in clustered data. This implies that adapting models for longitudinal data to other data structures is in many cases relatively straightforward. For example, clustered data could be analyzed by leaving out all aspects of the model that refer to time.

As in the univariate settings, a very important characteristic of data to be analyzed is the type of outcome. Methods for continuous data form the best developed and most advanced body of research; the same is true for software implementation. This is natural, since the special status and the elegant properties of the normal distribution simplify model building and ease software development. It is in this area that the general linear mixed model is situated. However, also categorical (nominal, ordinal, and binary) and discrete outcomes are very prominent in statistical practice. For example, quality of life outcomes are often scored on ordinal scales.

Two fairly different views can be adopted. The first one, supported by large-sample results, states that normal theory should be applied as much as possible, even to non-normal data such as ordinal scores and counts. A different view is that each type of outcome should be analyzed using instruments that exploit the nature of the data. In addition, since the statistical community has been familiarized with generalized linear models, some have taken the view that the normal model for continuous data is but one type of generalized linear models. Although this is correct in principle, it fails to acknowledge that normal models are much further developed than any other generalized linear models (e.g., model checks and diagnostic tools) and that it enjoys unique properties (e.g., the existence of closed-form solutions, exact distributions of test statistics, unbiased estimators, no mean-variance link, etc.). Extensions of generalized linear models to the longitudinal case include marginal models (e.g., generalized estimating equations) and random-effects models (e.g., the generalized linear mixed model).

In longitudinal settings, each individual typically has a *vector* $\boldsymbol{Y}$ of responses with a natural (time) ordering among the components. This leads to several, generally non-equivalent, extensions of univariate models. In a *marginal model*, marginal distributions are used to describe the outcome vector $\boldsymbol{Y}$, given a set X of predictor variables. The correlation among the components of $\boldsymbol{Y}$ can then be captured either by adopting a fully parametric approach or by means of working assumptions, such as in the generalized estimating equations approach. Alternatively, in a *random-effects model*, the predictor variables X are supplemented with a vector $\boldsymbol{b}$ of random effects, conditional upon which the components of $\boldsymbol{Y}$ are usually assumed to be independent. This does not preclude that more elaborate models are possible if residual dependence is detected. Finally, a *conditional model* describes the distribution of the components of $\boldsymbol{Y}$, conditional on X but also conditional on (a subset of) the other components of $\boldsymbol{Y}$. In a longitudinal context, a particularly relevant class of conditional models describes a component of $\boldsymbol{Y}$ given the ones recorded earlier in time. Well-known members of this class of *transition models* are *Markov type* models.

For normally distributed data, marginal models can easily be fitted with a number of standard statistical software packages. For such data, integrating a mixed-effects model over the random effects produces a marginal model, in which the regression parameters retain their meaning and the random effects contribute in a simple way to the variance-covariance structure. For example, the marginal model corresponding to a random-intercepts model is a compound-symmetry model that can be fitted without explicitly acknowledging the random-intercepts structure. In the same vein, certain types of transition model induce simple marginal covariance structures. For example, some first-order stationary autoregressive models imply an exponential or AR(1) covariance structure. As a consequence, many marginal models derived from random-effects and transition models can be fitted with mixed-models software.

It should be emphasized that the above elegant properties of normal models do not extend to the general non-Gaussian case. For example, opting for a marginal model for longitudinal binary data precludes the researcher from answering conditional and transitional questions in terms of simple model parameters.

While research in this area has largely focused on the formulation of linear mixed-effects models, inference, and software implementation, other important aspects, such as exploratory analysis, the investigation of model fit, and the construction of diagnostic tools have received considerably less attention. In addition, longitudinal data are typically very prone to incompleteness, due to dropout or intermediate missing values. This poses particular challenges to methodological development.

2.2.3 The Linear Mixed Model

The linear mixed-effects model is a commonly used tool for, among others, variance component models and for longitudinal data. We will briefly introduce it and discuss some of the issues surrounding this model. Let $\boldsymbol{Y_i}$ denote the n_i-dimensional vector of measurements available for subject $i = 1, \ldots, N$. A general linear mixed model then assumes that $\boldsymbol{Y_i}$ satisfies

$$\boldsymbol{Y_i} = X_i\boldsymbol{\beta} + Z_i\boldsymbol{b_i} + \boldsymbol{\varepsilon_i}, \tag{6}$$

in which $\boldsymbol{\beta}$ is a vector of population-average regression coefficients called fixed effects, and where $\boldsymbol{b_i}$ is a vector of subject-specific regression coefficients. The $\boldsymbol{b_i}$ describe how the evolution of the ith subject deviates from the average evolution in the population. The matrices X_i and Z_i are $(n_i \times p)$ and $(n_i \times q)$ matrices of known covariates. The random effects $\boldsymbol{b_i}$ and residual components $\boldsymbol{\varepsilon_i}$ are assumed to be independent with distributions $N(\boldsymbol{0}, D)$, and $N(\boldsymbol{0}, \Sigma_i)$, respectively. Inference for linear mixed models is usually based on maximum likelihood or restricted maximum likelihood estimation under the marginal model for $\boldsymbol{Y_i}$, i.e., the multivariate normal model with mean $X_i\boldsymbol{\beta}$, and covariance $V_i = Z_iDZ_i' + \Sigma_i$. Thus, we can adopt two *different* views on the linear mixed model. The fully *hierarchical* model is specified by

$$\begin{array}{rcl} \boldsymbol{Y_i}|\boldsymbol{b}_i & \sim & N_{n_i}(X_i\boldsymbol{\beta} + Z_i\boldsymbol{b}_i, \Sigma_i), \\ \boldsymbol{b}_i & \sim & N(0, D), \end{array} \tag{7}$$

while the marginal model is given by

$$\boldsymbol{Y_i} \sim N_{n_i}(X_i\boldsymbol{\beta}, Z_iDZ_i' + \Sigma_i). \tag{8}$$

Even though they are often treated as equivalent, there are important differences between the hierarchical and marginal views on the model. Making likelihood-based inferences on the marginal model, and deriving a satisfactory fit has been obtained, does not imply the posited hierarchical model is plausible. In fact, it can be shown that several hierarchical models can lead to the same marginal model. Moreover, some marginal models cannot be derived from a hierarchical model. A direct implication of this fact is that one has to be careful with inference on variance components in linear mixed models. A hierarchical interpretation of variance components implies that the null hypothesis associated to the tests lies on the boundary of the parameter space. This, in turn, implies that inference and in particular null distributions take a non-standard form.

Other issues, surrounding the linear mixed model, focus on data exploration and model confirmation. Exploring data and model building involves not only a mean structure and variance components, but the further decomposition of both in several constituents. For example, the mean structure typically has a time-independent or cross-sectional part, as well as a time-dependent or longitudinal part. The covariance structure typically includes, apart from the variance function, all three components of association alluded to earlier, i.e., random effects, serial association and measurement error (instantaneous replication error). In order to perform model criticism, it is important to realize that residuals can be defined in essentially two ways: conditional upon the random effects or marginalized over them. One or the other definition may make more or less sense, depending on the situation.

2.2.4 From Gaussian to Non-Gaussian Longitudinal Data

In the previous section we have discussed a number of issues arising from the use of the linear mixed effects model. Still focusing on continuous outcomes, a marginal model is characterized by the specification of a marginal mean function

$$E(Y_{ij}|\boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta}, \tag{9}$$

whereas in a random-effects model we focus on the expectation, conditional upon the random-effects vector:

$$E(Y_{ij}|\boldsymbol{b}_i, \boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \boldsymbol{z}_{ij}'\boldsymbol{b}_i. \tag{10}$$

Finally, a third family of models conditions a particular outcome on the other responses or a subset thereof. In particular, a simple first-order stationary transition model focuses on expectations of the form

$$E(Y_{ij}|Y_{i,j-1}, \dots, Y_{i1}, \boldsymbol{x}_{ij}) = \boldsymbol{x}_{ij}'\boldsymbol{\beta} + \alpha Y_{i,j-1}. \tag{11}$$

As we have seen before, random-effects models imply a simple marginal model in the linear mixed model case. This is due to the elegant properties of the multivariate normal distribution. In particular, the expectation (9) follows from (10) by either (a) marginalizing over the random effects or by (b) by conditioning upon the random-effects vector $\boldsymbol{b_i} = \boldsymbol{0}$. Hence, the fixed-effects parameters $\boldsymbol{\beta}$ have both a marginal as well as a hierarchical model interpretation. Finally, when a conditional model is expressed in terms of residuals rather than outcomes directly, it also leads to particular forms of the general linear mixed effects model.

Such a close connection between the model families does not exist when outcomes are of a non-Gaussian type, such as binary, categorical, or discrete outcomes. The main reason is that the left hand sides of (9)–(11) have to be replaced by $g[E(Y_{ij}|\cdot)]$ where g is an appropriate link function, thus destroying the simple links described above, since these exist thanks to the property of linearity. Let us therefore focus on some differences between the model families.

It will be clear from the briefest comparison of the marginal and the conditional families, that fitting a marginal model is typically more involved than fitting a conditional model. In addition, most marginal models have constrained parameter spaces. This is often quoted as an interpretational disadvantage. However, the same is true for the multivariate normal model since its covariance matrix has to be positive definite. In contrast, the parameters in many conditional models can take on any value in the Euclidean space whilst still producing valid probabilities. However, one of the main interpretational advantages of marginal models is their upward compatibility or reproducibility. This means that when a marginal model is used to model a response vector, then the appropriate sub-model applies to any subvector of the response vector. Precisely, such a sub-vector still follows a model of the same structure, with as parameter vector the corresponding sub-vector. This is not true in conditional models, posing particular problems when response vectors are of unequal length.

Marginal models should be chosen whenever there are marginal research questions, e.g., pertaining to one or a few occasions, or the evolution between them. They are also useful when not only the strength of association between occasions, but also a quantification of this association is of interest. Of course, when the number of measurement occasions within a subject grows, such models become intractable from a likelihood perspective. One can then resort to alternative approaches, such as generalized estimating equations or pseudo-likelihood. In fact, generalized estimating equations are by far the most commonly used marginal approach. This is a sensible approach when one is mainly interested in first-order marginal mean parameters. For clustered and repeated data, Liang and Zeger proposed so-called generalized estimating equations (GEE) which require only the correct specification of the univariate marginal distributions provided one is willing to adopt "working" assumptions about the association structure. They estimate the parameters associated with the expected value of an individual's vector of binary responses and phrase the working assumptions about the association between pairs of outcomes in terms of marginal correlations. The method combines estimating equations for the regression parameters $\boldsymbol{\beta}$ with moment-based estimating for the correlation parameters entering the working assumptions.

Models with subject-specific parameters are differentiated from population-averaged models by the inclusion of parameters which are specific to the cluster. Unlike for correlated Gaussian outcomes, the parameters of the random effects and population-averaged models for correlated binary data describe different types of effects of the covariates on the response probabilities.

The choice between population-averaged and random effects strategies should heavily depend on the scientific goals. Population-averaged models evaluate the overall risk as a function of covariates. With a subject-specific approach, the response rates are modeled as a function of covariates and parameters, specific to a subject. In such models, interpretation of fixed-effect parameters is conditional on a constant level of the random-effects parameter. Population-averaged comparisons, on the other hand, make no use of within cluster comparisons for cluster varying covariates and are therefore not useful to assess within-subject effects.

Whereas the linear mixed model is unequivocally the most popular choice in the case of normally distributed response variables, there are more options in the case of non normal outcomes. An early instance of a random-effects model for binary data is the beta-binomial model. The most popular model in this setting, however, is the generalized linear mixed model, where a generalized linear model (e.g., with logit link for binary data or log link for counts) is combined with a linear predictor that includes normally distributed random effects. While the idea seems natural, key issues arise. First, it should be clear that the parameters obtained from a marginal approach such as GEE and a GLMM approach are not directly comparable since they estimate different sets of population parameters. In the first case, the parameters describe an evolution over time or a covariate dependence over a population as a whole, while in the second case such effects are studied conditional upon a level of a subject's random effects. Second, fitting a GLMM requires integration of a subject's contribution to the log-likelihood function over the random-effects distribution. While this problem has a closed-form solution in the linear mixed-effects model, either numerical integration or an approximation to the integrand is required in the general cases.

2.3 Survival Analysis

Survival analysis refers to statistical procedures used to analyze data where the outcome of interest is time to an event (Kalbfleisch and Prentice 1980, Kleinbaum 1996). Examples of events include death and recurrence of illness. Survival data, or time-to-event data, occur in health sciences and elsewhere, such as in technometrics, actuarial sciences, etc. Even though there is a strong connection between the analysis of survival data and other types of data analysis, there is an important key difference in the sense that not all study subjects may experience the event. Such observations are called censored, the others being uncensored. Censored subjects are included since they do provide partial information, i.e., for them it is known in which time interval the event did not occur. It is commonly said that they contribute "follow-up time" to the study.

Let us first introduce some concepts and notation. Denote a survival time by a random variable T and its cumulative distribution function by $F(t)$, i.e., the probability that a failure occurs before time t. The corresponding density is denoted by $f(t)$. The range for t typically is $[0, +\infty[$. In addition, and this is specific to the survival area, one defines the survivorship function or survival function $S(t) = 1 - F(t)$, i.e., the probability tat the event occurs after time t. When the outcome is death, $S(t)$ indicates the probability to survive until time t. A very important concept is the hazard, $h(t)$, defined as the probability that an individual fails in a small interval of time conditional on their survival at the beginning of such interval. It may be written as:

$$h(t) = \frac{f(t)}{1 - F(t)}.$$

One often refers to it as the instantaneous failure rate or the force of mortality. The cumulative hazard function is then defined as

$$H(t) = \int_0^t h(x)\, dx.$$

Survival data are commonly represented using Kaplan-Meier plots. These are also considered a non-parametric way of analyzing such data. When a certain number of survival curves is to be compared (two or more), one often uses log-rank or Wilcoxon tests. These can be seen as the counterparts to analysis of variance and contingency table analysis in continuous and discrete data, respectively. Just as there are regression methods for continuous and categorical data (linear and logistic regression, respectively), a number of regression methods are available for (censored) time-to-event data. Fully parametric models include the exponential model

$$f(t) = \lambda \exp(-\lambda t)$$

for $t \geq 0$ and parameter $\lambda > 0$, and the Weibull model

$$f(t) = \lambda\gamma t^{\gamma-1} \exp(-\lambda t^{\gamma}),$$

for $t \geq 0$ and parameters $\lambda, \gamma > 0$. In the exponential model, the hazard is constant, i.e., $h(t) = \lambda$, whereas for the Weibull model it equals $h(t) = \lambda\gamma t^{\gamma-1}$. Now, λ can be modeled as a function of covariates, thus producing a flexible regression class for survival outcomes. This is especially true in the Weibull case, due to the presence of the shape parameter γ.

However, it is often considered a drawback having to model the entire distribution function, whereas interest focuses primarily on the difference between survival curves (e.g., between treated and untreated patients). This is why the Cox proportional hazards model has become so popular. Precisely, the hazard for the ith subject, $h_i(t)$ is written as

$$h_i(t) = h_0(t)\varphi(x_i),$$

where $h_0(t)$ is a so-called baseline hazard, common to all subjects and left unspecified, and $\varphi(x_i)$ specifies the effect of a set of covariates. This function must be nonnegative, explaining why often a log-linear specification is employed:

$$\ln \varphi(x_i) = \ln\left(\frac{h_i(t)}{h_0(t)}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1}.$$

This equation also reveals the assumption behind the model: the hazard for an individual is proportional to that for another individual, regardless of time. Alternate models are available to analyze data when this assumption is not satisfied, including the stratified Cox model. The stratified Cox model estimates regression coefficients for variables that do satisfy the proportional hazards assumption, stratifying on variables that do not satisfy the assumption. For the Cox model, conditional, permutation-based arguments are used to construct a so-called partial likelihood function. Specific attention is to be devoted to tied observations, and several methods for handling thereof exist. In addition, the method requires non-trivial but now standard extension when covariates vary with time. The Cox model, just as the fully parametric survival regression methods, requires specific attention with censored observations.

Just as with continuous and categorical outcomes, the occurrence of correlated survival times is becoming ever more frequent. For example, survival times may be observed on several members within the same family (e.g., in population genetics) or time to several non-exclusive events may be observed for the same person. The same split between marginal and random-effects models exist as in the settings considered earlier, but the methods are further complicated by the presence of censoring. An example of a marginal approach is provided by appropriately constructed generalized estimating equations. Random-effects models tend to be referred to as frailty models in this particular context. Special attention is devoted to so-called recurrent events, as well as to competing risks.

Recently, interest has increased in the joint modeling of longitudinal outcomes and survival times. This can be motivated from, at least, three different perspectives. First, the survival analyst can be interested in the effect of an entire longitudinal covariate, rather than merely in the effect of a time-varying covariate. Second, the longitudinal modeler can be left with a partially unobserved longitudinal profile due the operation of a time-to-event or dropout process. Third, one can be interested in the longitudinal and survival outcomes simultaneously, from a more symmetric perspective, in the sense that a longitudinal profile or a part thereof can be considered as a (candidate) surrogate marker or surrogate endpoint for a true endpoint that is of a time-to-event type.

In the following four sections, methodology specifically relevant for the area of EEG data will be presented.

3 Analysis of Variance (ANOVA)

Based on Neter *et al.* (1996), we will give an overview of the basics of ANOVA.

3.1 Notation

Let Y_{ij} represent the observation or measurement for subject $i = 1, \ldots, n_j$ for factor level (i.e., group) $j = 1, \ldots, r$. The total of the observations for the jth factor level is denoted by $Y_{\cdot j}$:

$$Y_{\cdot j} = \sum_{i=1}^{n_j} Y_{ij}.$$

Note that the dot in $Y_{\cdot j}$ indicates an aggregation over the i index, i.e., all subjects. The sample mean for the jth factor level is denoted by $\overline{Y}_{\cdot j}$:

$$\overline{Y}_{\cdot j} = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} = \frac{Y_{\cdot j}}{n_j}.$$

Finally, the overall mean for all responses is denoted by $\overline{Y}_{\cdot\cdot}$:

$$\overline{Y}_{\cdot\cdot} = \frac{\sum_{j=1}^{r} \sum_{i=1}^{n_j} Y_{ij}}{\sum_{j=1}^{r} n_j},$$

where the two dots indicate aggregation over all subjects i and all factor levels j.

3.2 Partitioning of the Variability

The total variability of the Y_{ij} observations, not using any information about factor levels, is measured in terms of the total squared deviation of each observation, i.e. the squared deviation of Y_{ij} around the overall mean $\overline{Y}_{..}$:

$$SSTO = \sum_{j=1}^{r} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{..})^2, \tag{12}$$

where $SSTO$ stands for *total sum of squares*.

When we utilize information about the factor levels, we can calculate the sum of squared deviations of all observation Y_{ij} around their respective estimated factor level means $\overline{Y}_{\cdot j}$:

$$SSE = \sum_{j=1}^{r} \sum_{i=1}^{n_j} (Y_{ij} - \overline{Y}_{\cdot j})^2, \tag{13}$$

where SSE stands for *error sum of squares*.

Finally, the *treatment sum of squares* is defined as:

$$SSTR = \sum_{j=1}^{r} \sum_{i=1}^{n_j} (\overline{Y}_{\cdot j} - \overline{Y}_{..})^2. \tag{14}$$

This brings us to a basic equality in ANOVA:

$$SSTO = SSTR + SSE. \tag{15}$$

In other words, the total variability can be decomposed as sum of the *between* and *within* group variability.

3.3 Testing for Treatment Differences

We will now present a test statistic using $SSTR$ and SSE which will enable us to test for treatment differences. This testing problem can be formulated as testing the following null hypothesis H_0 versus the alternative hypothesis H_a:

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \ldots = \mu_r, \\ H_a : \text{not all } \mu_j \text{ are equal,} \end{cases} \tag{16}$$

where the treatment means μ_j are the expected values for each response Y_{ij}.

More precisely, we use the following ANOVA model:

$$Y_{ij} = \mu_j + \varepsilon_{ij}. \tag{17}$$

The ε_{ij}'s are called *random errors* and they are assumed to be independent and normally distributed: $\varepsilon_{ij} \sim N(0, \sigma^2)$. A third critical assumption underlying this model is that the variance of the random errors remains constant across all factor levels j, the so-called *heterogeneity assumption*.

The test statistic to be used for choosing between the two hypotheses is the F-test:

$$F^* = \frac{MSTR}{MSE} = \frac{SSTR/(r-1)}{SSE/(n_T - r)} \sim F(r-1, n_T - r), \tag{18}$$

where n_T stands for the total number of subjects, MSE for *mean square error* and $MSTR$ for *mean square treatment*. Hence, for a given significance level α and with $F(1-\alpha; r-1, n_T - r)$ the $(1-\alpha)100$ percentile of the F distribution with $r-1$ and $n_T - r$ degrees of freedom, we have the following decision rule:

$$\begin{cases} \text{if } F^* \leq F(1-\alpha; r-1, n_T - r), \text{ conclude } H_0, \\ \text{if } F^* > F(1-\alpha; r-1, n_T - r), \text{ conclude } H_a, \end{cases}$$

where F^* is the actual value of the F-statistic. The so-called p-value associated with this F-statistic is the probability of observing the sample treatment differences under the assumption that the null hypothesis is true. A small p-value indicates that the null hypothesis is unlikely to be true.

4 Mixed-effects Models for Longitudinal Data

Based on the considerations in Section 1.3, we will now move to genuine longitudinal analysis techniques. After introducing the classical linear mixed model in Section 4.1, we shift to their non-linear extension in Section 4.2.

4.1 The Linear Mixed Models

The linear mixed-effects model (Laird and Ware 1982, Verbeke and Molenberghs 2000) is a commonly used tool for, among others, longitudinal data. It has been introduced in Section 2.2.3.

4.2 Non-linear Mixed Models

Perhaps the most commonly encountered subject-specific model is the generalized linear mixed model. Assume the data setting is the same as in Section 4.1. A general framework for mixed-effects models for longitudinal data can be expressed as follows. Assume that $\boldsymbol{Y_i}$ (possibly appropriately transformed) satisfies

$$\boldsymbol{Y_i} | \boldsymbol{b_i} \quad \sim \quad F_i(\boldsymbol{\theta}, \boldsymbol{b_i}), \tag{19}$$

i.e., conditional on $\boldsymbol{b_i}$, $\boldsymbol{Y_i}$ follows a pre-specified distribution F_i, possibly depending on covariate matrices X_i and Z_i (suppressed from notation), and parameterized through a vector $\boldsymbol{\theta}$ of unknown parameters, common to all subjects. Further, $\boldsymbol{b_i}$ is a q-dimensional vector of subject-specific parameters, called random effects, assumed to follow a so-called mixing distribution G which may depend on a vector $\boldsymbol{\psi}$ of unknown parameters, i.e., $\boldsymbol{b_i} \sim G(\boldsymbol{\psi})$. The $\boldsymbol{b_i}$ reflect the between-unit heterogeneity in the population with respect to the distribution of $\boldsymbol{Y_i}$. In the presence of random effects, conditional independence (upon $\boldsymbol{b_i}$) is often assumed,

In general, unless a fully Bayesian approach is followed, inference is based on the marginal model for $\boldsymbol{Y_i}$ which is obtained from integrating out the random effects, over their distribution $G(\boldsymbol{\psi})$ (Fahrmeir and Tutz 2001). Let $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$ and $g(\boldsymbol{b_i})$ denote the density functions corresponding to the distributions F_i and G, respectively, we have that the marginal density function of $\boldsymbol{Y_i}$ equals

$$f_i(\boldsymbol{y_i}) = \int f_i(\boldsymbol{y_i}|\boldsymbol{b_i})g(\boldsymbol{b_i})d\boldsymbol{b_i}, \tag{20}$$

which depends on the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. Assuming independence of the units, estimates of $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\psi}}$ can be obtained from maximizing the likelihood function built from (20), and inferences immediately follow from classical maximum likelihood theory.

It is important to realize that the random-effects distribution G is crucial in the calculation of the marginal model (20). One approach is to leave G unspecified and to use non-parametric maximum likelihood (NPML, McLachlan and Peel 2000) estimation, which maximizes the likelihood over all possible distributions G. The resulting estimate $\widehat{G}$ is then always discrete with finite support. Depending on the context, this may or may not be a realistic reflection of the true heterogeneity between units. One therefore often assumes G to be of a parametric form, such as a (multivariate) normal. Depending on F_i and G, the integration in (20) may or may not be possible analytically. Proposed solutions are based on Taylor series expansions of $f_i(\boldsymbol{y_i}|\boldsymbol{b_i})$, or on numerical approximations of the integral, such as (adaptive) Gaussian quadrature (Pinheiro and Bates 1995).

Although one is usually primarily interested in estimating the parameters in the marginal model, it is often needed to calculate estimates for the random effects $\boldsymbol{b_i}$ as well, e.g., for predictive purposes or to detect special profiles, outlying individuals, or groups of individuals evolving differently in time. Inference for the random effects is often based on their posterior distribution $f_i(\boldsymbol{b_i}|\boldsymbol{y_i})$, given by

$$f_i(\boldsymbol{b_i}|\boldsymbol{y_i}) \quad = \quad \frac{f_i(\boldsymbol{y_i}|\boldsymbol{b_i})\ g(\boldsymbol{b_i})}{\int\ f_i(\boldsymbol{y_i}|\boldsymbol{b_i})\ g(\boldsymbol{b_i})\ d\boldsymbol{b_i}}, \tag{21}$$

in which the unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are replaced by estimates obtained from maximizing the marginal likelihood. The mean or mode corresponding to (21) can be used as point estimates for $\boldsymbol{b_i}$, yielding empirical Bayes (EB) estimates.

There are two major differences in comparison to the linear mixed model. First, the marginal distribution of $\boldsymbol{Y_i}$ can no longer be calculated analytically, such that numerical approximations to the marginal density (8) come into play, complicating the computation of the MLE for $\boldsymbol{\beta}$, D, and the parameters in all Σ_i. As a result, the marginal covariance structure does not immediately follow, such that it is not always clear in practice what assumptions a specific model implies with respect to the underlying variance function and the underlying correlation structure in the data.

A second difference is related to the interpretation of the fixed effects $\boldsymbol{\beta}$. Under the linear model (6), $\mathsf{E}(\boldsymbol{Y_i}) = X_i\boldsymbol{\beta}$, such that the fixed effects have a subject-specific as well as a marginal interpretation: the elements in $\boldsymbol{\beta}$ reflect the effect of specific covariates, conditionally on $\boldsymbol{b_i}$, as well as marginalized over these random effects. Under non-linear mixed models, this does not generally hold. The fixed effects now only reflect the conditional effect of covariates, and the marginal effect is not easily obtained anymore as $\mathsf{E}(\boldsymbol{Y_i})$ is given by

$$\mathsf{E}(\boldsymbol{Y_i}) \quad = \quad \int \boldsymbol{y_i} \int f_i(\boldsymbol{y_i}|\boldsymbol{b_i})g(\boldsymbol{b_i})d\boldsymbol{b_i}d\boldsymbol{y_i},$$

which, in general, is *not* of the form $h(X_i, Z_i, \boldsymbol{\beta}, \mathbf{0})$.

Only for very particular models, (some of) the fixed effects can still be interpreted as marginal covariate effects. For example, consider the model where, apart from an exponential link function, the mean is linear in the covariates, and the only random effects in the model are intercepts. More specifically, this corresponds to the model with $h(X_i, Z_i, \boldsymbol{\beta}, \boldsymbol{b_i}) = \exp(X_i\boldsymbol{\beta} + Z_i b_i)$, in which Z_i is now a vector containing only ones. The expectation of $\boldsymbol{Y_i}$ is now given by

$$\mathsf{E}(\boldsymbol{Y_i}) = \mathsf{E}\left[\exp(X_i\boldsymbol{\beta} + Z_i b_i)\right] = \exp(X_i\boldsymbol{\beta})\ E\left[\exp(Z_i b_i)\right], \qquad (22)$$

which shows that, except for the intercept, all parameters in $\boldsymbol{\beta}$ have a marginal interpretation.

5 Discrimination and Classification

A very specific problem is the classification of drugs. Drug classifications can be based on a variety of different considerations and there appears to be little general agreement as to the optimal scheme for ordering the universe of biologically active substances. For example, drugs might be organized according to chemical structure, clinical-therapeutic use, potential health hazards, liability to non-medical use, public availability and legality, effects on specific neural or other physiological systems, or influence on certain psychological and behavioural processes. The classification systems developed from these different approaches may show considerable overlap, although there are often striking incongruities. For example, some drugs which appear very similar in chemical structure may be quite different in pharmacological activity and vice-versa. The most useful organization depends on the intended use of the classifications.

There are different approaches to classification. First, it can be done intuitively. For example, a physician or a group of physicians may use their experience in caring for patients with chest pain to form a subjective opinion or an empirical decision as to whether a new patient with chest pain is likely to suffer a heart attack, and consequently, decide what treatment is most appropriate. Secondly, methods in both statistical and machine learning literature have been developed, such as Fisher linear discriminant analysis (Fisher 1936). These methods have the parametric flavor in the sense that the classification rule has an explicit form with only a few parameters to be determined from a given sample that is usually referred to as learning sample. Classification trees belong to the third type of methods for which we allow a very general structure, e.g., the binary tree, but the number of "parameters" also needs to be determined from the data, and this number varies. For this reason, classification trees are regarded as nonparametric methods. They are adaptive to the data and are flexible, although the large number of quantities (or parameters) to be estimated from the data makes the classification rule more vulnerable to noise in the data.

We will focus our attention on the classification of psychotropic drugs based on the changes induced by the drugs in sleep-waking behaviour using electroencephalographic (EEG) spectral. Discrimination and classification belongs to the family of multivariate methods, as introduced in Section 2.2.1. Section 6 describes a non-parametric technique, classification and regression trees, which can be used for classification purposes. The methods are exemplified in Section 8.

5.1 Discriminant Analysis

In many cases, the subgroup (stratified) structure of the data, will be the focus of scientific interest. Two very distinct situations can arise:

Known groups: Groups have been defined explicitly. This is often based on subject matter (e.g., biological) knowledge. The drug classes have been identified and the classification is widely accepted by the scientific community. Of course, this does not imply that it is easy to *discriminate* between groups.

Unknown groups: The researcher has good grounds to believe in the existence of strata, even though they have not been defined clearly. She might even be uncertain about the actual *number of groups*.

The second situation is the topic of **cluster analysis**. In this section, attention is confined to **discriminant analysis**. An EEG data set will be used as an illustration. This is in line with tradition, as part of discriminant analysis is called **Fisher's discriminant analysis**.

Discriminant function analysis is used to determine which variables discriminate between two or more naturally occurring groups. In linear discriminant analysis (LDA) we are interested in those linear features which reduce the dimensionality and simultaneously preserve class separability. Discriminant analysis can be understood as an exploratory tool to describe the dependence relations of the response variable on the given set of predictors in the observed sample of cases; the G categories of the response variable define a partition of the population Ω into G groups $(\omega_1, \omega_2, ..., \omega_G)$ and the P predictors are observed to characterize the typologies of cases within each group (Saporta 1990, McLachlan 1992).

At the same time, discriminant analysis can also be used to define a decision rule for assigning a new case to one class on the basis of the observations of the given predictors in the so-called learning sample; a method such as test sample or cross-validation is considered to estimate the accuracy of the decision rule (Fisher and van Ness 1973, Celeux and Nakache 1994).

To summarize the discussion so far, the basic idea underlying discriminant function analysis is to determine whether groups differ with regard to the mean of a variable, and then to use that variable to predict group membership (e.g., of new cases).

Discriminant analysis can serve two slightly different purposes.

Discrimination: Suppose the five drug classes have been clearly defined. The question is then whether it is possible to distinguish among them in an "optimal" way. Indeed, when several characteristics are recorded, it might turn out that summary measures, such as mean and standard deviation, are distinct among the drug classes.

Again, one can try and pick the variable with the "largest" differences among group (e.g., by means of an ANOVA model). However, this would waste information on all other variables. As before, the idea is to combine information on the available variables, e.g., by incorporating all of them into a linear combination.

This can be viewed as an exploratory, descriptive technique. It can also be used for a graphical display of the relative location of the drug classes.

Classification: Given a sample for which group membership is known, and given a new observation, the question is to allocate the new compound to a particular population.

Clearly, this is not a descriptive technique any more. A (mathematical) rule is required, an automated decision process, which indicated unambiguously to which drug classes a new compound should be allocated, given the values on the characteristics.

In *both* cases, one often relies on an algebraic rule, a **discriminant function**.

5.2 The Scope of Discriminant Analysis

Discriminant analysis has been studied extensively, and can be approached from several distinct angles.

5.2.1 Fully Parametric or Not

In regression analysis, distinction can be made between a semi-parametric (moment based) approach, such as *least squares*, and a fully parametric *likelihood* approach. The distinction is academic, since point estimators under both paradigms coincide, while standard precision measures are asymptotically the same.

In a similar fashion, discriminant analysis can be approached by two philosophically different roads:

The parametric way: This method is based on assuming a parametric distributional form for the outcomes in each of the subgroups(drug classes). Differences between these distributions (in terms of their parameters), are used to discriminate between them.

The best known examples include normal and logistic discriminant analysis. We will focus on normal discriminant analysis. A further subdivision can be made between:

- a normal distribution in each group, with different means (location parameters), but with the same variance-covariance matrix.
- a normal distribution in each group, with both mean vectors and covariance matrix different.

Fisher's way: This method is concerned with finding a linear combination of the original variables, that displays the group differences best. (Of course, "best" will have to be defined properly).

Just as in regression analysis, where least squares (a "linear" technique), and likelihood (a "normal distribution" technique) yield the same estimators, both normal discriminant analysis (with equal variances) and Fisher's (linear) discriminant analysis are in fact two faces of the same coin.

5.2.2 Two or More Subgroups

As soon as there are at least two subgroups, the discriminant analysis problem is legitimate. While the theory for more groups is in principle a fairly straightforward extension of the theory for two groups, the latter one is easier in notation. Therefore, emphasis will be placed on the two groups case first.

5.2.3 Quality of the Classification Rule

It is not because a classification rule has been determined that it is of any practical relevance, even not if the rule turns out to be "statistically significant" in a well defined sense. It might happen that two groups have been defined very clearly while several characteristics of the subjects in these groups tend to be similar. In other words, the **group means** could be close to each other, relative to the random variability of the subjects within each of the groups separately. This situation arises when the two clouds of points are poorly separated.

In statistical terms: the **between group variability** is too small, compared to the **within group variability**, to lead to a useful separation criterion.

How can this then yield a *statistically significant* difference ? Consider a single variable X, measured for a number of subjects in two subgroups, 1 and 2. Suppose the group means are 20 and 21 respectively, and that the standard deviation in each of the groups is 5. Clearly, the two clouds of points are going to show a considerable overlap. Intuitively, we feel the classification boundary should be 20.5. When the sample size in each group increases, the group means are going to be estimated more and more precisely, and as a consequence, a significant difference will be found eventually. However, by applying this classification rule, a lot of **errors** are going to be made, no matter how large the sample size.

Thus, for small samples, errors occur for two reasons:

- the classification rule is imprecise (small sample size; incorrect distributional assumptions);
- the populations genuinely overlap.

For (very) large samples, errors for the first reason could disappear, but the population overlap is a population characteristic, not a sampling characteristic, and will not disappear... Therefore, it is important to evaluate the classification rule. This leads to the study of **classification error**. A few concerns about classification errors:

- Some groups are much less frequent than others (e.g., a very rare disease versus the healthy subpopulation). This information should be included.
- When we have two groups, two classification errors can be made. A subject that truly belongs to subgroup 1 can be classified into group 2 and vice versa. However, often one error is much less severe than another. For instance, in a breast cancer screening program, a false positive test is much less dramatic than a false negative, since the first one allows correction during a more thorough investigation, while for the second one the tumor can start to develop.
- Using the same sample for constructing a classification rule as well as to test it might be misleading (too optimistic). Indeed, while the classification rule is constructed to learn something about the population as a whole, it will often perform better on the sample it was constructed on than on most other samples. This is true because the rule will adapt to features in the sample that happened purely by chance (e.g., outliers), while it will not cope explicitly with such features in new samples.

  This fact should be accounted for when developing methods to evaluate a classification rule.

  Standard procedures are: the use of learn+validation samples, cross-validation, and the bootstrap.

5.2.4 Purpose of Classification

We present a few situations where a classification criterion is useful.

- For each subject, measurements are made at different times:

  - predictors at time 0,
  - outcomes at time 1.

  Given a set of complete measures, can I predict/classify the outcome for new individuals, using only their predictor information?

- Given non-destructive predictors and destructive outcomes, can I, using a learn sample, determine a rule to classify an object w.r.t. the outcomes, only using the non-destructive predictors?

- Some variables are unavailable for all objects, so we want to use predictor variables.

- Outcome too expensive.

A prediction rule, using easy, generally available measures, is used to predict outcomes of complicated measures. This is based on a training sample which contains both sets of measures.

5.3 Parametric Version: Two Populations

For the parametric theory, we will restrict attention to two groups (two drug classes, two classes of objects). Indicate the two classes of objects by π_1 and π_2. Recall that we want to:

- distinguish between them,

- allocate objects to them.

For each object or individual i, a set of p measures $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})^T$ is obtained (e.g., the time spent in each of the sleeping stages). We hope that the measurements are "different" between the two groups (e.g., the mean of some of them is higher or lower in one group than in the other).

A difference between populations is translated into statistical language by the claim that they are generated by a different stochastic mechanism, which in turn is characterized by a different distribution. An observation in population $j = 1, 2$ follows distribution $F_j(\boldsymbol{x})$ with density $f_j(\boldsymbol{x})$.

Remarks:

- Almost always, we are not at all sure about our classification. This is translated into a probabilistic statement: we claim that a (new) compound belongs to population 1 with probability 0.7 (posterior probability).

- Rephrase the question as: "Given the vector of measurements $\boldsymbol{x}$, does it come from π_1 or from π_2?"

We have introduced the notation π_1 and π_2 to indicate the two groups. At the same time, the observation vector $\boldsymbol{x}$ occupies a value in its space. For example, the active and quiet wake variables will be recorded in the following space:

$$\Omega = [0, +\infty[\times[0, +\infty[\times[0, +\infty[\times[0, +\infty[.$$

Since we want to operate at the level of our observed measurements, the problem can be further reduced to the following aim:

> Divide the variable space Ω into 2 parts R_1 and R_2 (regions) and adopt the rule:
>
> A new observations is assumed to belong to π_j if it falls in R_j.

Of course, the regions should be a partition:

- The union of R_1 and R_2 fill the whole parameter space Ω.
- The intersection of R_1 and R_2 is empty (has probability zero).

5.3.1 Classification Error

Classification is bound to err. In statistical terms, we have to study the **misclassification error**:

- X_i belongs to π_1 and is classified into π_2,
- X_i belongs to π_2 and is classified into π_1.

Thus, based on the classification rule and the observations made for a particular compound, we are lead to believe that the compound belongs to one subgroup, whereas in reality it belongs to the other.

It is important to realize that for many data configurations, a perfect classification (error free) is not possible. This will often be clear from a simple **graphical inspection** of the data.

5.3.2 Properties of a Good Classification Rule

The following properties, discussed before, should be sought for a good classification rule:

1. the *misclassification probabilities* are minimal,
2. the *prior probabilities* are taken into account:

- if one population is much larger than another, classification in the largest should be more frequent;
- e.g. there are more antidepressant than stimulants drugs, whence classification of a compound as stimulants candidate should occur only if the evidence is overwhelming.

3. *cost* of misclassification error (ethical cost, economic cost),

    - e.g. classifying a healthy person as diseased implies further investigation and eventually the healthy condition of the person will be established, while the opposite misclassification is dramatic. Note that similar remarks apply in controlling type I and type II error rate. We sometimes need to distinguish between the two classification errors.

5.3.3 Formalizing the Classification Error

Recall the setting:

- Two populations π_1 and π_2 with associated densities $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$.
- Let Ω be the sample space and suppose it is partitioned as $\Omega = R_1 \cup R_2$ with R_j the set of $\boldsymbol{x}$ which we would classify as belonging to π_j (which could be wrong).

Our previous statements about classification that will often never be perfect can be formalized as follows: Note that in general, there is an *optimal* classification possible, but no *perfect* classification.

The **classification errors** are:

- the probability of lying in the second region and belonging to the first population:

$$P(2|1) = P(\boldsymbol{X} \in R_2|\pi_1) = \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x}.$$

- The opposite classification error is:

$$P(1|2) = P(\boldsymbol{X} \in R_1|\pi_2) = \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}.$$

Now, consider the **prior probabilities**:

- $p_1 =$ prior probability of belonging to π_1,
- $p_2 =$ prior probability of belonging to π_2.

Evidently $p_1 + p_2 = 1.$

For example, we might have an idea about the relative proportions of the five drug classes.

There are 4 "classification probabilities". A compound can belong to π_1 or π_2 and can be classified as population π_1 and π_2, leading to a 2×2 factorial:

		classify as	
		π_1	π_2
true	π_1	$P(1\mid 1)$	$P(2\mid 1)$
population	π_2	$P(1\mid 2)$	$P(2\mid 2)$

With a slight abuse of notation, the probabilities of correct and incorrect classification can be written in terms of the prior probabilities p_1 and p_2, and of the quantities $P(j|k)$.

The correct classification probability for population π_1:

$$\begin{aligned} P(\text{correctly classified as } \pi_1) &= P(\boldsymbol{X} \in \pi_1, \boldsymbol{X} \in R_1) \\ &= P(\boldsymbol{X} \in \pi_1).P(\boldsymbol{X} \in R_1 | \boldsymbol{X} \in \pi_1) \\ &= p_1.P(1|1), \end{aligned}$$

and the misclassification error is

$$\begin{aligned} P(\text{misclassified as } \pi_1) &= P(\boldsymbol{X} \in \pi_2, \boldsymbol{X} \in R_1) \\ &= P(\boldsymbol{X} \in \pi_2).P(\boldsymbol{X} \in R_1 | \boldsymbol{X} \in \pi_2) \\ &= p_2.P(1|2). \end{aligned}$$

In summary:

$$\begin{aligned} P(\text{correctly classified as } \pi_1) &= p_1.P(1|1), \\ P(\text{misclassified as } \pi_1) &= p_2.P(1|2), \\ P(\text{correctly classified as } \pi_2) &= p_2.P(2|2), \\ P(\text{misclassified as } \pi_2) &= p_1.P(2|1). \end{aligned}$$

These probabilities are the first step to answer the following questions:

1. What is the misclassification error ?

(2.) What is the misclassification cost ?

The **cost matrix** is very simple in the case of two groups:

		classify as	
		R_1	R_2
true	π_1	0	$c(2\mid 1)$
population	π_2	$c(1\mid 2)$	0

We are now able to compute the

5.3.4 Expected Cost of Misclassification

$$\begin{aligned}\mathsf{ECM} &= c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2\\ &= c(2|1)p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + c(1|2)p_2 \int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}\\ &= c(2|1)p_1 \int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + c(1|2)p_2 \left(1 - \int_{R_2} f_2(\boldsymbol{x})d\boldsymbol{x}\right)\\ &= c(1|2)p_2 + \int_{R_2} \big\{f_1(\boldsymbol{x})[c(2|1)p_1] - f_2(\boldsymbol{x})[c(1|2)p_2]\big\}d\boldsymbol{x}.\end{aligned}$$

Minimizing ECM is done by choosing those points that yield a negative contribution to the integral:

$$\begin{aligned}R_2 &= \{\boldsymbol{x}|f_1(\boldsymbol{x})c(2|1)p_1 - f_2(\boldsymbol{x})c(1|2)p_2 < 0\}\\ R_2 &= \{\boldsymbol{x}|f_1(\boldsymbol{x})c(2|1)p_1 < f_2(\boldsymbol{x})c(1|2)p_2\}\\ R_2 &= \left\{\boldsymbol{x}\left|\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} < \frac{c(1|2)p_2}{c(2|1)p_1}\right.\right\}.\end{aligned}$$

Similarly,

$$R_1 = \left\{\boldsymbol{x}\left|\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{c(1|2)p_2}{c(2|1)p_1}\right.\right\}.$$

and hence the regions are defined.

What if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{c(1|2)p_2}{c(2|1)p_1}\ ?$$

This boundary case is fairly arbitrary and the performance of the rule will not change when we either assign this curve to R_1 or to R_2.

5.3.5 Structure of the ECM

The classification rule is: Assign an observation with outcome vector x to the first population π_1 if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{c(1|2)p_2}{c(2|1)p_1}.$$

In other words, the ratio of the densities should exceed a threshold function.

The boundary:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \frac{c(1|2)p_2}{c(2|1)p_1}$$

is called the **discriminant function**.

Inspecting this ratio, it is clear that we only need:

prior probability ratio p_1/p_2. Of course, knowing the ratio is equivalent to knowing p_1 only or to knowing p_2 only. Indeed, the quantities sum to one, whence they represent only 1 independent quantity.

cost ratio $c(1|2)/c(2|1)$. This is an important reduction of the information that needs to be found. Even if the components are hard to specify, the ratio can be much easier to establish. Indeed, one might have difficulty in calculating even a rough approximation of the actual cost involved in these misclassifications. But it is plausible that one has a rough idea about the relative severity of the misclassification, e.g., the second type of misclassification is 10 times as bad as one of the first type.

A few remarks are in place.

- The shape of the discriminant function depends on the forms of $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$. It will change with changing parametric forms assumed for these densities (e.g., normal densities with equal or with unequal variances).
- If either the cost ratio or the prior probability ratio is unity, the definition of the regions simplifies accordingly.
- If the product of cost and prior probability ratio is unity, then we actually allocate to the population with the highest probability. We then classify to R_1 if

  $$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > 1,$$

  or, equivalently,

  $$f_1(\boldsymbol{x}) > f_2(\boldsymbol{x}).$$

The ECM is not the only useful criterion to determine the classification boundary. A few alternatives are:

Total probability of misclassification (TPM) : the ECM for equal costs.

Largest posterior probability : reduces to the TPM.

5.4 Two Multivariate Normal Populations

Now that we have defined a classification criterion:

Assign an observation with outcome vector $\boldsymbol{x}$ to the first population π_1 if

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} > \frac{c(1|2)p_2}{c(2|1)p_1},$$

we can focus on a few standard cases.

Assume a multivariate normal form for the two populations:

$$\begin{array}{lcl} \pi_1 & : & N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), \\ \pi_2 & : & N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \end{array}$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are the mean vectors, and $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are the covariance matrix. Here $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$ and, $\boldsymbol{\Sigma}_2$ are assumed to be known.

The underlying principle is that the mean vector shifts in switching from one population to the other. This feature is then used to "draw a line" (a plane, a hyperplane,...) between the two population. We need to distinguish between the situation where the covariance matrices are equal or unequal.

5.4.1 Equal Covariance Matrices

In this case, we assume $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}$.

Explicitly, the densities are $(i = 1, 2)$:

$$f_i(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_i)\right].$$

The classification rule is based on the ratio of the two densities, evaluated at $\boldsymbol{x}$:

$$\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})} = \exp\left[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_1) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_2)\right].$$

After some manipulations, the classification region R_1 is found to be:

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq \ln\left[\frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1}\right].$$

Sample Version In the above reasoning, $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$ are assumed to be known population values. However, in practice, they are unknown.

This implies they have to be estimated from data. The following algorithm can be used:

- Collect n_1 observations out of π_1 and n_2 observations out of π_2.
- Construct the sample statistics $\overline{\boldsymbol{x}}_1$, $\overline{\boldsymbol{x}}_2$, $\boldsymbol{S}_1$, and $\boldsymbol{S}_2$, as estimators for $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$, respectively.
- Since we assume a common $\boldsymbol{\Sigma}$, it is necessary to construct a common $\boldsymbol{S}$. In other words, $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$ are assumed to estimate the same quantity, and therefore, they should be combined, in a so-called **pooled sample covariance matrix**:

$$\boldsymbol{S}_{\text{pooled}} = \frac{(n_1 - 1)\boldsymbol{S}_1 + (n_2 - 1)\boldsymbol{S}_2}{(n_1 + n_2 - 2)}.$$

> Observe that, when the sample sizes n_1 and n_2 are equal, then $\boldsymbol{S}_{\text{pooled}}$ is simply the average of $\boldsymbol{S}_1$ and $\boldsymbol{S}_2$, otherwise, they are weighted by the sample size they are based upon.

Plugging in these estimators leads to the

Estimated Minimum ECM Rule for 2 Normal Populations Allocate an observation with measurements $\boldsymbol{x}_0$ to π_1 if

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_{\text{pooled}}^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geq \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right].$$

If the product of the two ratios is unity, then

$$\ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right] = 0,$$

and the right hand side of the allocation rule vanishes, whence it can be rewritten as

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_{\text{pooled}}^{-1} \boldsymbol{x}_0 \geq \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_{\text{pooled}}^{-1}(\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2).$$

Remark. We did not explicitly prescribe a particular parametric form of the *classification rule*. Rather, the form of the densities involved was specified a priori. Yet, a very simple, **linear discriminant function** arises. Once again, normal theory implies linear theory. (Here, linearity is respect to $\boldsymbol{x}_0$).

Define the linear combination vector

$$\boldsymbol{\ell}^T = (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}_{\text{pooled}}^{-1}.$$

This linear combination occurs both on the left hand side, as well as on the right hand side of the classification rule.

The rule can be rewritten as:

$$\boldsymbol{\ell}^T \boldsymbol{x}_0 \geq \frac{1}{2}(\boldsymbol{\ell}^T \overline{\boldsymbol{x}}_1 + \boldsymbol{\ell}^T \overline{\boldsymbol{x}}_2) = \boldsymbol{m}.$$

Some remarks are in place.

- $\boldsymbol{\ell}$ is called the vector of **discriminant coefficients**.
- Our rule is only an estimate of the optimal rule, we do not know the population versions $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, and $\boldsymbol{\Sigma}$. This implies that
  - the sample size should be reasonably large ($n_i - p \geq 20$),
  - normality must hold in each sub-population,
  - the two covariance matrices must be equal.

By rewriting the rule, we clearly see that a **univariate** variable $\boldsymbol{\ell}^T \boldsymbol{x}_0$ is compared to the average of the two univariate means $\boldsymbol{\ell}^T \overline{\boldsymbol{x}}_1$ and $\boldsymbol{\ell}^T \overline{\boldsymbol{x}}_2$, i.e. the "midpoint" $\boldsymbol{m}$.

Should we have had a univariate observation for each compound (e.g., active wake only, rather than 6 outcome variables), then the midpoint of one of the drug class mean and another drug class mean would be the natural candidate as a "classification rule". Since we were confronted with a multivariate setting, it was not a priori clear what a classification rule (discriminant function) would look like. However, we are back to the univariate setting, since we have constructed a linear combination, which reduces the 6 original variables to a single new one, to which we then apply our simplistic first idea of computing the midpoint. Of course, this linear combination is **optimal** in the sense that it minimizes the Expected Cost of Misclassification.

A warning is in place. This nice and simple result is found by the virtue of three assumptions:

- both populations have a normal dispersion;
- the covariance matrices are equal;
- the product of the prior probability ratio and the cost ratio is unity.

Should the ratio be different from unity, then the rule is slightly more complex, since it has to make the "most expensive" misclassification less likely.

The contribution of each of the original variables is determined by their coefficients ℓ_j. The question that one wants to answer is:

How important are the discriminators $X_1, \dots, X_p$?

- Define
$$\boldsymbol{\ell}^{(1)} = \frac{\boldsymbol{\ell}}{\sqrt{\boldsymbol{\ell}^T \boldsymbol{\ell}}},$$
yielding a vector with unit length, of which all components lie within $[-1, 1]$.

  Their magnitudes can be interpreted, *if the original variables have been standardized.*
- Alternatively, define
$$\boldsymbol{\ell}^{(2)} = \frac{\boldsymbol{\ell}}{\ell_1}$$
enabling to compare $X_2, \dots, X_p$ with X_1, as the coefficient of X_1 is 1.

5.4.2 Unequal Covariance Matrices

We now allow that $\boldsymbol{\Sigma}_1 \neq \boldsymbol{\Sigma}_2$.

Manipulating the ratio of the densities, R_1 is defined as the set of vectors satisfying:

$$R_1 : -\frac{1}{2}\boldsymbol{x}^T(\boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\Sigma}_2^{-1})\boldsymbol{x} + \left(\boldsymbol{\mu}_1^T \boldsymbol{\Sigma}_1^{-1} - \boldsymbol{\mu}_2 \boldsymbol{\Sigma}_2^{-1}\right)\boldsymbol{x} - k \geq \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right],$$

where

$$k = \frac{1}{2}\ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_2|}\right) + \frac{1}{2}\left(\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2^T\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\mu}_2\right).$$

If $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$ then the quadratic term vanishes and we again obtain the linear discriminant function.

Plugging in the sample versions we obtain the quadratic classification rule:

Quadratic Classification Rule for Two Normal Populations Allocate $\boldsymbol{x}_0$ to π_1 if

$$-\frac{1}{2}\boldsymbol{x}_0^T(\boldsymbol{S}_1^{-1} - \boldsymbol{S}_2^{-1})\boldsymbol{x}_0 + \left(\overline{\boldsymbol{x}}_1^T\boldsymbol{S}_1^{-1} - \overline{\boldsymbol{x}}_2\boldsymbol{S}_2^{-1}\right)\boldsymbol{x}_0 - k \geq \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right].$$

Guidelines:

- If the populations are approximately normal and the variances are unequal: use the quadratic classification rule.
- BUT: the quadratic rule is sensitive to departures from normality, while the linear rule is much more generally valid, **also outside the normal framework**, as we will learn from Fisher's discriminant analysis.
- carry out checks before performing a classification procedure:
  - transform to normality first
  - then check for homogeneity of the covariance matrix

  The order is important since these homogeneity checks are sensitive to nonnormality.

5.5 Evaluating Classification Functions

As mentioned earlier, it is important to assess the performance of a classification rule. This performance is a function of two arguments:

- How precise is the actual discriminant function an estimate of the true discriminant function? This is largely due to sampling variability and can be improved by increasing the learning sample size.
- How well are the two populations separated? If they are different but largely overlapping, any discriminant function, no matter how precisely determined, will be bound to err a lot.

5.5.1 Derivation of Optimum Error Rate

Recall that the Total Probability of Misclassification (TPM) is

$$\text{TPM} = \text{TPM}(R_1, R_2) = p_1\int_{R_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2\int_{R_1} f_2(\boldsymbol{x})d\boldsymbol{x}.$$

Clearly, the TPM is a function of the regions R_1 and R_2, and the problem has been to choose the regions carefully, in order to minimize the TPM.

The regions for which the minimum is achieved is called the **optimum error rate** (OER):

$$\text{OER} = \min_{R_1,R_2}(TPM).$$

For example, the OER in the case of two normal populations with the same shape, equal priors $p_1 = p_2 = 0.5$ and equal costs is obtained for the regions:

$$\begin{aligned} R_1 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) \geq 0 \\ R_2 &: (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) < 0. \end{aligned}$$

In this case, the OER can be computed explicitly, and turns out to be:

$$\text{OER} = \Phi\left(-\frac{\Delta}{2}\right),$$

with

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2),$$

and $\Phi(.)$ the standard normal cumulative density function.

This result is intuitively appealing:

- The larger Δ, the smaller $-\Delta/2$, and thus the smaller the OER.
- Now, Δ increases when:
  - $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$ increases, i.e., when the means of the two groups are further apart and thus better separated. In other words, when the **between group variability increases**.
  - $\boldsymbol{\Sigma}$ decreases, i.e., when the **within group variability decreases**.

Thus, the ideal situation is given by two compact clouds of points that are far apart.

5.5.2 Actual Error Rate

Above computation can be carried out if the populations are completely known. This is of course a theoretical situation and in practice, the sample version of the parameters have to be used. In this case, the terminology changes to **actual error rate**, formally defined as:

$$\text{AER} = p_1\int_{\widehat{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x} + p_2\int_{\widehat{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x},$$

with

$$\begin{aligned}
\widehat{R}_1 &: (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) \geq \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right], \\
\widehat{R}_2 &: (\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x} - \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2) < \ln\left[\frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}\right].
\end{aligned}$$

Problem:

- $f_1(\boldsymbol{x})$ and $f_2(\boldsymbol{x})$ unknown,
- we will have to find an estimated version.

Although work has been done to carry out this estimation, and results are even implemented in SAS PROC DISCRIM, this falls outside of the scope of these notes. We will discuss a simpler alternative.

5.5.3 Apparent Error Rate

A generally applicable error rate, not requiring the parent distribution, is the **apparent error rate** (APER): "What fraction in the training (learning) sample is misclassified ?" We first construct the so-called *confusion matrix*:

		predicted		
		π_1	π_2	
actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

Then,

$$\text{APER} = \frac{n_{1M} + n_{2M}}{n_1 + n_2} = \frac{(n_1 + n_2) - (n_{1C} + n_{2C})}{n_1 + n_2}.$$

- **Advantage**: this quantity is easy to compute.
- **Disadvantage**: too optimistic (underestimates AER). This implies that the evaluation done at the learning sample, will be typically much better than the performance for test samples.

  Why ? The rule is ready-made for the learn sample, and this very sample is used to evaluate the rule.
- **Solution**:

  Construct LEARN+TEST sample (TRAINING+VALIDATION sample):

  - **Advantage**: this overcomes bias,

- **Disadvantage**: requires larger samples and/or considerable computation time.

Alternative: **cross-validation technique**:

- cycle through all observations $i = 1, \dots, n_1 + n_2$,
- construct classification rule without observation i,
- classify observation i.

We then get an unbiased estimate of the expected actual error rate:

$$\widehat{E}(\mathsf{AER}) = \frac{n_{1M}^{(J)} + n_{2M}^{(J)}}{n_1 + n_2}.$$

In the next section we will briefly describe a non-parametric technique that can also be used for classification purposes.

6 Classification Tree Analysis

Classification tree analysis is a method where, following specific splitting rules, disjoint subsets of the data are constructed. These subsets are called nodes. Further splitting is repeated several times within these nodes. A node where a split is formed is called a parent node; the subsequent nodes are called child nodes. Terminal nodes are nodes that are not split further. The size of the tree is the number of parent nodes plus one. We focus on binary classification trees, where splitting occurs into exactly two child nodes. This partitioning process results in a saturated tree. A tree is saturated in the sense that the offspring nodes subject to further division cannot be split. The saturated binary tree is then pruned to an optimal size tree. This is the so-called pruning process. The final step is the selection process, which determines the final tree. In the following sections a brief overview of the different processes is given.

6.1 The Partitioning Process

The partitioning process is based on splitting rules. The splitting rules involve conditioning on predictor variables. The best possible variable to split the root node is the one that results in the most homogeneous and purest child nodes. A measure for the goodness of split is defined as the reduction in impurity. The best split is the split with the largest reduction in impurity. The splits are selected one at a time, starting with the split at the root node (including all individuals), and continuing with splits of resulting child nodes until splitting stops, and the child nodes that have not been split become terminal nodes. This partitioning process results in a saturated tree with the characteristic that if no limit is placed on the number of splits that are performed, eventually ‘pure’ classification (all subjects have the same value with reference to the dependent variable) will be achieved. However, ‘pure’ classification is usually unrealistic. The saturated tree is usually too large to be useful, the terminal nodes are so small that no sensible inference can be made, so the tree has a small predictive value. Therefore it is typically to set a minimum size of a node a priori or a maximum number of levels for the tree to reach (Breiman *et al.* 1984).

6.2 The Pruning Process

The point is to find the subtree of the saturated tree that is most predictive of the outcome and least vulnerable to noise in the data. Breiman *et al.* (1984) proposed to let the partitioning continue until the tree is saturated or nearly so, and this generally large tree is pruned from the bottom up. The method of cost-complexity pruning is used. This function is defined as the cost for the tree plus a complexity parameter times the tree size. The cost of a tree is a measure for total impurity in the final nodes. The sum of the impurities in the terminal nodes is indicative for the quality of the tree. The larger the tree the smaller the impurities in the terminal nodes, but the more complex the tree is. The tree size is a measure of the tree complexity. The procedure generates a sequence of trees which are nested and optimally pruned, because for every size of a tree in the sequence, there is no other tree of the same size with lower costs.

6.3 The Selection Process

For the original dataset, the cost decreases monotonically with increasing number of nodes. This corresponds to the fact that the maximum tree will give the best fit. For the test data, the cost decreases with increasing number of nodes, but reaches a minimum and then increases as complexity increases. This reflects that an overfitted and overly complex tree will not perform well on new data. The optimal tree corresponds to the complexity parameter that gives a minimum cost for the new data. Often there are several trees with costs close to the minimum, then the smallest-sized tree whose cost does not exceed the minimum cost plus 1 times the standard error of the cost will be chosen. This is the ‘1 SE rule’. When no test sample is available, V-fold cross-validation is useful. A specified V value for V-fold cross-validation determines the number of random subsamples, as equal in size as possible, that is formed from the learning sample. The classification tree of the specified size is computed V times, each time leaving out one of the subsamples from the computations, and using that subsample as a test sample for cross-validation. The CV costs computed for each of the V test samples are then averaged to give the V-fold estimate of the CV costs.

6.4 Missing Data

One attractive feature of tree-based methods is the ease with which missing values can be handled. There are several methods to deal with missing values. The method used here was the approach of surrogate splits, which attempt to utilize information in the other

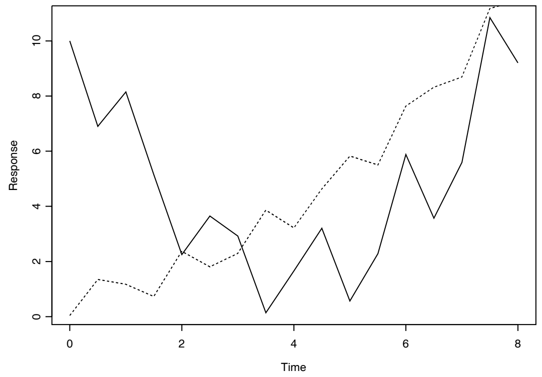

Figure 1: Mean profiles over time for treated (solid) and placebo (dotted) group.

predictors to assist in making the decision to send an observation to the left or to the right daughter node. They look for the predictor that is most similar to the original predictor in classifying the observations. Similarity is measured by a measure of association. It is not unlikely that the predictor that yields the best surrogate split may also be missing. Then there will be looked for the second best, and so on. In this way all available information is used.

7 Application of ANOVA and Longitudinal Methods

In this section, we will present an application of statistical methodology for pharmaco-EEG data. First, we will describe the experimental design and consecutively explore the data from a descriptive point of view. This is followed by a discussion of the results of a linear mixed-effects model analysis. Finally, the results of a non-linear mixed-effects approach are presented.

7.1 Experimental Design

In this experiment 60 subjects are randomly assigned to either the placebo or the treatment group. The subjects were monitored over a period of 8 hours and the time spent in a certain sleeping stage was summarized every 30 minutes.

7.2 Data Exploration

The mean profiles over time for the treated and placebo group are shown in Figure 1. We can see that the longitudinal profiles are clearly different between both groups. The observed profile for the placebo group is almost monotonically increasing over time, while the profile is roughly U-shaped for the treated group.

7.3 ANOVA Model

We can try to simplify the data by summarizing the longitudinal profile into one summary statistic per subject, which in this case is the sum of all observed values over time. The results of analyzing this summary statistic are presented in the following ANOVA table:

```
                                       Sum of
Source                      DF        Squares     Mean Square    F Value    Pr > F

Model                        1      12.411713       12.411713       0.08    0.7786

Error                       58    9026.006643      155.620804

Corrected Total             59    9038.418356
```

There appears to be no significant difference between the placebo and the treated group (p=0.7786).

Collapsing the data into one summary statistic might not be the optimal solution for this experiment. Another approach is to analyze the treatment difference by time point. This approach will be discussed next.

7.4 Analysis By Time Point

An overview of the estimated treatment differences and ANOVA per time point is presented Table 1. If we want to maintain an overall significance level α of 5% by applying the Bonferroni correction, the adjusted critical p-value becomes:

$$\frac{\alpha}{2 \times g} = \frac{0.05}{2 \times 17} = 0.00147,$$

where g stands for the number of comparisons performed (Neter *et al.* 1996), i.e., the number of time points.

By comparing this adjusted p-value to the list presented in Table 1, we observe that there is a significant difference between both groups at time point 1, 2, 3, 4, 8, 11, 12, 14 and 15. The estimated difference is negative for the first four significant time points, indicating that the placebo group has lower observed response values compared to the treatment group. The other time points associated with significant p-values have positive estimates, indicating higher response values.

7.5 Linear Mixed Model

The following linear mixed-effects model was fitted to these data:

$$Y_{ij} = \beta_0 + b_i + \beta_1 \text{ Time} + \beta_2 \text{ Time}^2 + \beta_3 \text{ Group} + \beta_4 \text{ Time*Group} + \beta_5 \text{ Time}^2\text{*Group} + \varepsilon_{ij}. \tag{23}$$

Table 1: Estimated treatment difference and ANOVA per time point.

Time point	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
T1	-8.0155	0.8555	986	-9.37	<.0001
T2	-5.0719	0.8555	986	-5.93	<.0001
T3	-5.7704	0.8555	986	-6.74	<.0001
T4	-3.3895	0.8555	986	-3.96	<.0001
T5	0.8634	0.8555	986	1.01	0.3131
T6	-2.3710	0.8555	986	-2.77	0.0057
T7	-0.7760	0.8555	986	-0.91	0.3646
T8	2.8450	0.8555	986	3.33	0.0009
T9	1.7434	0.8555	986	2.04	0.0418
T10	2.3333	0.8555	986	2.73	0.0065
T11	3.5376	0.8555	986	4.13	<.0001
T12	3.2572	0.8555	986	3.81	0.0001
T13	1.6010	0.8555	986	1.87	0.0616
T14	5.2333	0.8555	986	6.12	<.0001
T15	3.3279	0.8555	986	3.89	0.0001
T16	0.1177	0.8555	986	0.14	0.8906
T17	1.4441	0.8555	986	1.69	0.0917

This type of model is called a *random intercept* model, since the only subject-specific effect is the intercept b_i. The interaction terms Time*Group and Time^2*Group allow for a different linear and quadratic trend over time for both groups.

Fitting model (23) in SAS leads to the following output:

```
                      The Mixed Procedure

                   Solution for Fixed Effects

                                 Standard
Effect          Estimate            Error      DF    t Value    Pr > |t|

Intercept         1.6482           0.5150      58       3.20      0.0022
time            -0.00269           0.2634     956      -0.01      0.9919
time2             0.1349          0.02845     956       4.74      <.0001
group            10.0666           0.7283     956      13.82      <.0001
time*group       -4.2456           0.3725     956     -11.40      <.0001
time2*group       0.3423          0.04023     956       8.51      <.0001
```

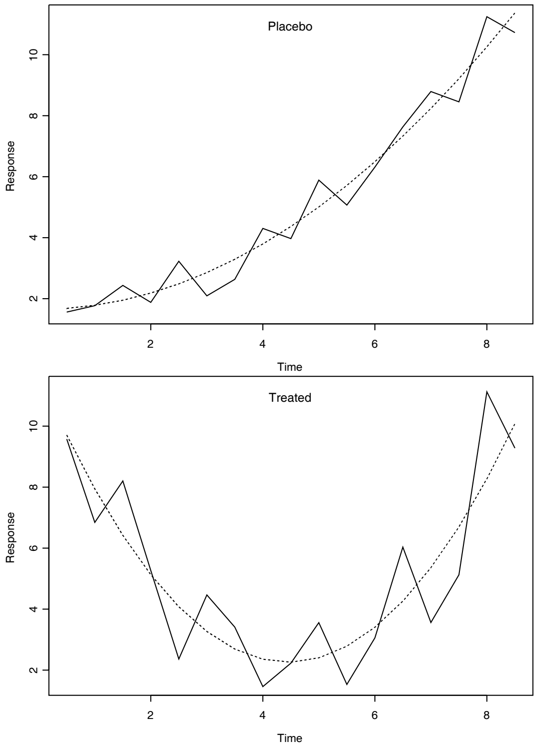

Figure 2: Mean (solid) and fitted (dotted) profiles over time for treated and placebo group.

From this output we can see that the estimated response $\widehat{Y}_{ij}$ for the placebo group equals

$$\widehat{Y}_{ij} = 1.6482 - 0.00269 \times \text{Time} + 0.1349 \times \text{Time}^2.$$

For the treatment group, the estimated response $\widehat{Y}_{ij}$ becomes

$$\widehat{Y}_{ij} = 11.7148 - 4.24829 \times \text{Time} + 0.4772 \times \text{Time}^2.$$

The observed and fitted longitudinal profiles for model (23) are presented in Figure 2. In this figure shows that the fitted curves seem to describe the data reasonably well for both groups. The differences in time and time^2 effect are highly significant.

Looking at the SAS output and Figure 2, we can clearly see that model (23) is able to capture the differences in the longitudinal pattern between both groups.

7.6 Non-linear Mixed Model

Fitting a non-linear mixed-effects model to this particular dataset is not needed, since the linear model had a satisfactory fit.

7.7 Some Reflections on Longitudinal Data Analysis

The analysis by time point with Bonferroni correction post-hoc has some serious drawbacks. If a large number of hypothesis tests are performed (which is often the case in EEG data), this will have a negative impact on the power. Furthermore, this approach usually leads to a large amount of p-values due to the high time resolution of the EEG data and drawing a clear and simple conclusion from such a list of p-values is not always evident.

Other multiple comparison procedures exist besides the Bonferroni correction. Especially the Dunnett correction can be of interest if all dose groups are compared with a control (i.e., placebo). Although a gain in power can be achieved in this setting by applying the Dunnett correction instead of the Bonferroni, the overall power of an analysis by time point is usually not satisfactory in pharmaco-EEG studies.

The summary statistic approach discussed in Section 1.4 might work for certain settings where the longitudinal profiles are roughly stable over time and the profiles do not cross, but as already mentioned, this is not the case in EEG data and therefore this approach is also not recommendable. For example, applying a summary statistic analysis to the data presented in Figure 1 did not enable us to find any difference between the treated and non-treated group, since e.g. the sum of the response values per subject is approximately the same for both groups. By taking into account the trend over time, e.g. as we did in Section 7.5 by fitting a linear mixed-effects model, we were able to find a treatment effect.

8 Application of Discrimination and Classification Methods

Two types of analysis will be discussed in this section. A parametric analysis (linear discriminant analysis) based on the time spent for each rat on a particular sleeping stage during both period (light and dark) was conducted. The second analysis was based on classification trees, the same variables were used, in order to be able to compare the results. The analysis were carried out using PROC DISCRIM (SAS procedure) for the classical discriminant analysis and RPART (SPLUS function) as far as classification and regression trees are concerned.

8.1 Discriminant analysis

8.1.1 Class Information

The following output of the procedure shows the structure of the data in use and the number of degrees of freedom. The between classes degrees of freedom is related to the group means, in this case we have 6 classes or groups. The within classes degrees of freedom is associated to the residual variability, and it is just the degrees of freedom associated to the overall mean minus the degrees of freedom associated to the group means. It is also important to remark that in each group the prior probability are chosen to be 1/6, but in case that this is in contradiction with the general knowledge, other prior probabilities can be used.

```
                     Normal Discriminant Analysis
                      Linear Classification Rule

                        Discriminant Analysis

         Observations     257          DF Total               256
         Variables         12          DF Within Classes      251
         Classes            6          DF Between Classes       5


                         Class Level Information

           Variable                                                  Prior
class      Name        Frequency       Weight    Proportion    Probability

antidep    antidep            31      31.0000      0.120623       0.166667
antipsy    antipsy            16      16.0000      0.062257       0.166667
anxiol     anxiol              8       8.0000      0.031128       0.166667
hypnot     hypnot             20      20.0000      0.077821       0.166667
placebo    placebo           150     150.0000      0.583658       0.166667
stimul     stimul             32      32.0000      0.124514       0.166667
```

8.1.2 Association Structure

The association structure can be represented by different measures:

- by means of sums of squares and cross-products (SSCP) matrices;
- by means of covariance matrices;
- by means of correlation matrices.

Since the dataset is structured into different subgroups, each of these measures has several versions:

- by subgroup: a separate matrix for each drug class;

- a pooled within subgroup matrix (the weighted average of the subgroup specific matrices);
- the total matrix (computed from all observations, ignoring the subgroup structure);
- the between subgroup matrix (the "difference" between the total and the within subgroup structure).

For the SSCP matrix, the computations are as follows:

by subgroup: the observations are corrected for the group mean, squared, and summed;

pooled within: the six "by subgroup" matrices are summed, resulting in the "pooled within" matrix;

between: the six group mean vectors are corrected for the grand mean, squared, and summed;

total: the observations are corrected for the grand mean, squared, and summed; this matrix equals the sum of within and between matrices.

From the SSCP matrices, the respective covariance matrices and correlation matrices can be derived.

Once this association structures are computed we can compare the *between* and *within* structure. In case that the between SSCP is much larger than the within SSCP, it implies that at least some separation should be possible. This can be also confirm using the covariance matrices. The last ones are more relevant to interpret since they are corrected for the correct number of degrees of freedom. When a discriminant analysis is used and the within class correlation structure is very different from the between correlation structure it should be in line with the results derived from the principal components analysis. In principal component analysis, the relevant choice is the within structure. When discriminant analysis is performed, both structures need to be contrasted.

8.1.3 Mean Structure (Location Structure)

In addition to the association structure already studied, PROC DISCRIM provides information on the location structure.

First, the grand mean structure is given. This consists of the average of each of the 12 variables, for each of the 258 observations in the data. This is augmented with the drug class specific averages.

```
                         Discriminant Analysis     Simple Statistics

                                       Total-Sample

Variable        N           Sum          Mean      Variance    Deviation

TAW_min1      257         40000     155.64362          3996      63.2112
TQW_min1      257         13411      52.18237          5042      71.0037
TSWS1_min1    257         47452     184.63646          4511      67.1631
TSWS2_min1    257         37435     145.66039          3550      59.5793
TIS_min1      257          1434       5.58043      14.40643       3.7956
TPS_min1      257         12531      48.76039     456.02702      21.3548
TAW_min2      257         54865     213.48374          2275      47.6982
TQW_min2      257          8081      31.44447          1046      32.3446
TSWS1_min2    257         11885      46.24708     745.98976      27.3128
TSWS2_min2    257         10153      39.50588     396.56354      19.9139
TIS_min2      257     480.72000       1.87051       2.37708       1.5418
TPS_min2      257          4422      17.20432     110.39708      10.5070
```

Somewhat less straightforward is the next panel of output:

```
                        Total-Sample Standardized Class Means

Variable         antidep      antipsy       anxiol       hypnot      placebo       stimul

TAW_min1    -0.183094469 -0.052826784  1.303786334 -0.088040092 -0.254189480  1.124377821
TQW_min1     1.133561132 -0.130871767 -0.439007321 -0.373542380 -0.067040309 -0.375234194
TSWS1_min1  -0.139328802  0.632906127 -1.210136229 -0.131120565  0.072972759 -0.139053683
TSWS2_min1  -0.378239119 -0.396652159  0.472577784  0.387418195  0.118732665 -0.352094958
TIS_min1    -1.108872567  0.364785302  0.508044020  0.989590912  0.191390658 -0.750821384
TPS_min1    -1.382770674 -0.237845434  0.020586056  0.320542190  0.373709031 -0.498764658
TAW_min2    -0.813829950 -0.307977552  0.851813811  0.358719036  0.164651754 -0.266571407
TQW_min2     1.269994102 -0.391606227 -0.310081818 -0.274480830 -0.079541586 -0.412581516
TSWS1_min2   0.320670088  0.163857830 -0.823361975 -0.160129303 -0.127181397  0.509506042
TSWS2_min2  -0.164469530  0.418601988 -0.337810493 -0.324038707 -0.020160562  0.331508314
TIS_min2    -0.411667471  1.180128071 -0.095186171  0.452396308 -0.114222648  0.085206340
TPS_min2    -0.187935792  0.957819027 -0.365881653  0.142112946 -0.197365449  0.630953650

                       Pooled Within-Class Standardized Class Means

Variable         antidep      antipsy       anxiol       hypnot      placebo       stimul

TAW_min1    -0.209883433 -0.060555989  1.494546248 -0.100921436 -0.291380515  1.288888071
TQW_min1     1.250109667 -0.144327515 -0.484144420 -0.411948617 -0.073933144 -0.413814377
TSWS1_min1  -0.143837117  0.653385312 -1.249293067 -0.135363283  0.075333966 -0.143553096
TSWS2_min1  -0.388287092 -0.407189276  0.485131876  0.397710011  0.121886814 -0.361448408
TIS_min1    -1.345144245  0.442511488  0.616294884  1.200446795  0.232170990 -0.910801739
TPS_min1    -1.706231130 -0.293482710  0.025401588  0.395524055  0.461127788 -0.615436675
TAW_min2    -0.870800747 -0.329537003  0.911443603  0.383830559  0.176177923 -0.285232290
TQW_min2     1.441321418 -0.444435483 -0.351913104 -0.311509399 -0.090272066 -0.468240423
TSWS1_min2   0.330905077  0.169087764 -0.849641636 -0.165240231 -0.131240710  0.525768204
TSWS2_min2  -0.166211241  0.423034927 -0.341387860 -0.327470234 -0.020374060  0.335018943
TIS_min2    -0.437624070  1.254537913 -0.101187882  0.480920956 -0.121424654  0.090578799
TPS_min2    -0.200663321  1.022685168 -0.390660165  0.151737226 -0.210731581  0.673683568
```

These are means for each of the subgroups, corrected for the grand mean and standardized. There are two distinct ways to standardize:

- compute the averages for each subgroup;
- subtract the overall mean;
- divide by:
  - the total sample standard deviation (i.e., the standard deviation computed from all 258 observations, ignoring the subgroup structure), leading to the **total sample standardized class means**;
  - the pooled within subgroup standard deviation (i.e., a combination of the standard deviations, calculated for each subgroup), leading to the **pooled within class standardized class means**.

We can make the following observations. Since the means are centered (subtraction of grand mean), the row totals are all zero. When the total sample standardized class means are smaller, relative to the pooled within class standardized class means, then the total sample standard deviations are much larger than the within class standard deviations. The difference between both is that the "total" version is the sum of "within" and "between", implying in turn that the between group differences are important.

8.1.4 The Discriminant Function

Up to now, the output has been given with the intention to describe and summarize the data. This process is independent of the actual classification rule adopted, whether it is normal or not, and whether it is linear or quadratic.

From now on, the discriminant function itself, as well as supporting quantities and derived results, are central. Therefore, we will see differences between the *linear* and *quadratic* versions. For each output panel, these two versions will be contrasted.

Rank and Conditioning of Covariance Matrix The discriminant functions and the distances between subgroup involve inverses of covariance matrices. For example,

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$$

for two populations and a linear (pooled) rule. It is important that the covariance matrices involved have full rank (determinant different from zero) and are not ill-conditioned (determinant sufficiently different from zero). This is expressed in the following outputs.

Linear Version The relevant quantity is the pooled covariance matrix.

```
  Linear Classification Rule

Pooled Covariance Matrix Information

                Natural Log of the
  Covariance    Determinant of the
 Matrix Rank     Covariance Matrix

          12              59.38441
```

Quadratic Version A different covariance matrix is used for different subgroups.

```
     Quadratic Classification Rule

                          Natural Log of the
            Covariance    Determinant of the
class      Matrix Rank     Covariance Matrix

antidep             12              52.73326
antipsy             12              45.81615
anxiol               7             -26.58118
hypnot              12              48.32568
placebo             12              58.03033
stimul              12              52.13322
```

8.1.5 Distances Between Groups

We have indicated that the quality of a classification rule depends on the separation between groups (together with the appropriateness of the rule itself). Once again, for two populations in the linear case, this distance is

$$\Delta^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$

It measures the distance in **relative terms**, i.e., the within group variability is taken into account. This is often referred to as the **Mahalanobis distance**.

Linear Version The following output is produced.

```
                               Linear Classification

                         Pairwise Squared Distances Between Groups

                              2         _   _       -1  _   _
                             D (i|j) = (X - X )' COV   (X - X )
                                         i   j           i   j


                                 Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep            0      11.03938       9.74957       9.39979       7.06291       6.97914
antipsy     11.03938             0      11.59851       6.49696       5.85131       9.49521
anxiol       9.74957      11.59851             0       7.61226       6.19096       4.79231
hypnot       9.39979       6.49696       7.61226             0       3.05915       9.97237
placebo      7.06291       5.85131       6.19096       3.05915             0       6.15873
stimul       6.97914       9.49521       4.79231       9.97237       6.15873             0


                F Statistics, NDF=12, DDF=240 for Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep            0       9.28292       4.94002       9.10534      14.45821       8.75647
antipsy      9.28292             0       4.92898       4.60165       6.74082       8.07030
anxiol       4.94002       4.92898             0       3.46603       3.74661       2.44389
hypnot       9.10534       4.60165       3.46603             0       4.30159       9.77983
placebo     14.45821       6.74082       3.74661       4.30159             0      12.94247
stimul       8.75647       8.07030       2.44389       9.77983      12.94247             0


                 Prob > Mahalanobis Distance for Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep       1.0000        <.0001        <.0001        <.0001        <.0001        <.0001
antipsy       <.0001        1.0000        <.0001        <.0001        <.0001        <.0001
anxiol        <.0001        <.0001        1.0000        <.0001        <.0001        0.0051
hypnot        <.0001        <.0001        <.0001        1.0000        <.0001        <.0001
placebo       <.0001        <.0001        <.0001        <.0001        1.0000        <.0001
stimul        <.0001        <.0001        0.0051        <.0001        <.0001        1.0000


                   Pairwise Generalized Squared Distances Between Groups

                       2         _   _       -1  _   _
                      D (i|j) = (X - X )' COV   (X - X ) - 2 ln PRIOR
                                  i   j           i   j              j


                          Generalized Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep      4.23018      16.59236      16.68884      14.50648       8.13980      11.14582
antipsy     15.26956       5.55297      18.53778      11.60365       6.92819      13.66189
anxiol      13.97975      17.15149       6.93927      12.71895       7.26784       8.95899
hypnot      13.62997      12.04993      14.55153       5.10669       4.13603      14.13905
placebo     11.29309      11.40428      13.13023       8.16584       1.07688      10.32541
stimul      11.20931      15.04818      11.73158      15.07906       7.23561       4.16668
```

At the beginning of above panel, the quantity $D^2(i|j)$ is the generalization of Δ^2 to the several subgroups case. At the end of the panel, the same function and the same distance matrix is given, but now labeled **pairwise generalized squared distance between groups**. The relevance will be understood in the light of the quadratic rule.

Clearly, the resulting distance matrix is symmetric, in the sense that the distance between drug classes antidep and antipsy is equal to the distance between drug classes antipsy and antidep. Of course, this property seems logical.

In addition to the actual distances, an F test and associated p values are computed for each of the distances. This is very natural, since Mahalanobis distances *always* lead to F tests. These distances $D^2(i|j)$ involve vectors and matrices, but:

- the **between group matrix**, $(\overline{X}_i - \overline{X}_j)(\overline{X}_i - \overline{X}_j)^T$ plays the role of the **numerator**, with 12 degrees of freedom since there are 12 variables.
- the **within group matrix**, $\boldsymbol{S}$ plays the role of the **denominator**, with 240 degrees of freedom.

In conclusion, the separation between any two drug classes, on the basis of the 12 outcomes, is highly significant.

Quadratic Version The quadratic version differs in several respects from the linear version:

- The quantity $D^2(i|j)$ is now defined **asymmetrically** ! Since the covariance matrices are unequal, and the covariance structure defines the (relative) coordinate system, this asymmetry is natural. For example, the covariance matrix of drug class antidep is apparantly "larger" than the covariance matrix of antipsy (although the concept "larger" is not evident in a matrix setting):
  - Antidep seems to be fairly far away (4617) from antipsy looking from inside the cloud of points of antipsy.
  - Looking from inside the cloud of antidep, the other cloud antipsy looks relatively close (40.04).

  In particular, the distance is defined as

  $$D(i|j)^2 = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).$$
- No F tests are given.
- The **generalized** version, correcting for the difference in covariance matrix, is now different from the standard version.

```
                           Quadratic Classification Rule

                      Pairwise Squared Distances Between Groups

                        2             _   _         -1  _   _
                       D (i|j) = (X - X )' COV   (X - X )
                                  i   j       j     i   j


                              Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep             0          4617     214098211     235.40224       7.00828      77.80888
antipsy      40.04586             0      86038960      29.44388       8.52089      17.74053
anxiol       43.54243      33.10573             0      11.27420       9.76251      24.19259
hypnot       46.32953     145.81507      34679175             0       3.38098      30.11011
placebo      23.13658     115.68214      97078087      14.53231             0       6.76926
stimul       13.76129      36.76037      19733684      31.13745      11.70143             0

                  Pairwise Generalized Squared Distances Between Groups

                  2             _   _         -1  _   _
                 D (i|j) = (X - X )' COV   (X - X ) + ln |COV | - 2 ln PRIOR
                            i   j       j     i   j          j             j


                         Generalized Squared Distance to class

From
class         antidep       antipsy        anxiol        hypnot       placebo        stimul

antidep      56.96344          4669     214098191     288.83460      66.11549     134.10878
antipsy      97.00930      51.36912      86038940      82.87624      67.62810      74.04043
anxiol      100.50587      84.47485     -19.64191      64.70656      68.86971      80.49249
hypnot      103.29298     197.18420      34679155      53.43237      62.48819      86.41001
placebo      80.10002     167.05126      97078067      67.96468      59.10721      63.06916
stimul       70.72473      88.12949      19733664      84.56981      70.80864      56.29990
```

Is Separation Possible? In the parametric discriminant theory, the rule is to set up a discriminant function, with the ability to allocate new observations to one of the subgroups. Repeatedly, we have emphasized that the quality of a rule depends on essentially two (sets of) properties:

Discriminant Function Dependent: Is the optimal discriminant function well approximated (sample size large enough? Are the correct distributional assumptions made, e.g., are they normal or not? If they are normal, is a linear or a quadratic rule in place?

Population Structure Dependent: Are the subgroups well separated? Do we see a clear difference between the clouds of points or do they overlap too much?

The next panel will address the second question, independent of the first one. Of course, this question has to be addressed using the *sample* at hand, since the theoretical population is out of reach.

Recall that we always (implicitly) assume that the observations are identified with the vector of (12) observations that are made for each compound. Thus, a poor separation could be improved by including extra variables. Alternatively, each outcome separately could be assessed as a “separator”.

Therefore,

- the next panel presents univariate and multivariate summaries of resolution power;
- it is independent of our choice for the linear or the quadratic version.

```
                             Univariate Test Statistics

                               Univariate Test Statistics

                         F Statistics,    Num DF=5,   Den DF=251

                         Total     Pooled    Between
                      Standard   Standard   Standard             R-Square
Variable    Label     Deviation  Deviation  Deviation  R-Square  / (1-RSq)  F Value  Pr > F

TAW_min1    AW min      63.2112    55.1431    34.8196    0.2538     0.3402    17.08  <.0001
TQW_min1    QW min      71.0037    64.3839    34.1770    0.1938     0.2404    12.07  <.0001
TSWS1_min1  SWS1 min    67.1631    65.0580    20.7731    0.0800     0.0870     4.37  0.0008
TSWS2_min1  SWS2 min    59.5793    58.0375    17.1872    0.0696     0.0748     3.76  0.0027
TIS_min1    IS min       3.7956     3.1289     2.3972    0.3337     0.5009    25.14  <.0001
TPS_min1    PS min      21.3548    17.3064    13.9312    0.3560     0.5529    27.76  <.0001
TAW_min2    AW min      47.6982    44.5777    19.7635    0.1436     0.1677     8.42  <.0001
TQW_min2    QW min      32.3446    28.4999    17.2797    0.2388     0.3137    15.75  <.0001
TSWS1_min2  SWS1 min    27.3128    26.4680     8.4062    0.0792     0.0861     4.32  0.0009
TSWS2_min2  SWS2 min    19.9139    19.7052     4.3529    0.0400     0.0416     2.09  0.0672
TIS_min2    IS min       1.5418     1.4503     0.6133    0.1324     0.1526     7.66  <.0001
TPS_min2    PS min      10.5070     9.8406     4.2976    0.1400     0.1627     8.17  <.0001


                                    Average R-Square

                            Unweighted              0.1717542
                            Weighted by Variance    0.1550737


                      Multivariate Statistics and F Approximations

                                   S=5    M=3    N=119

      Statistic                        Value    F Value    Num DF    Den DF    Pr > F

      Wilks' Lambda               0.20092182       7.68        60    1127.6    <.0001
      Pillai's Trace              1.30503451       7.18        60      1220    <.0001
      Hotelling-Lawley Trace      2.03002881       8.07        60    808.66    <.0001
      Roy's Greatest Root         0.91933493      18.69        12       244    <.0001

               NOTE: F Statistic for Roy's Greatest Root is an upper bound.
```

Let us concentrate on the results:

- Univariately, all variables but TSWS2_min2 are found to be highly significant predictors ($p \leq 0.01$). The values of the F statistics allow us to order them from highest to lowest resolution.

- The multivariate statistics are once again a comparison of between (group means) and within (pooled within covariance) structure. Therefore, the same *eigenvalue statistics*, used with canonical correlation analysis, can be used again.

  Again, they clearly indicate the significant discrimination between subgroups.

  Since there are $g = 6$ drug classes and $p = 12$ variables, the numerator degrees of freedom are $(g-1)p = 60$ for the first three and $p = 12$ for the last one.

  As an aside, let us discuss Hotelling-Lawley and Roy in some detail. The first one is

  $$\text{Tr}(\boldsymbol{W}^{-1}\boldsymbol{B}) = \sum_{i=1}^{2} \lambda_i = \lambda_1 + \lambda_2,$$

  whereas Roy is simply λ_1. Now, from the individual variables we see that TPS_min1 (total time spent in the light period in paradoxical sleeping stage) is the best separator and thus λ_1 might well favor this variable and downplay the others. Hence, the F value for TPS_min1 and Roy is very close. On the other hand, Hotelling-Lawley is very close to the average of all 12 F values.

  Thus, when there is one clear direction (approximately TPS_min1) in which differences are seen, with the others merely nuisances, Roy is able to produce a higher F values on less degrees of freedom. In other words, in this case, Roy is more powerful than the others (including Hotelling-Lawley), that look in all directions at once.

  Of course, examples of the reverse abound as well.

8.1.6 The Classification Rule

Linear Version The linear classification rule was given by

$$(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x}_0 \geq \frac{1}{2}(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2)^T \boldsymbol{S}^{-1}_{\text{pooled}} (\overline{\boldsymbol{x}}_1 + \overline{\boldsymbol{x}}_2).$$

Rewriting this rule (for populations i and j) gives

$$-\frac{1}{2}\overline{\boldsymbol{x}}_i^T \boldsymbol{S}^{-1}_{\text{pooled}} \overline{\boldsymbol{x}}_i + \overline{\boldsymbol{x}}_i^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x}_0 \geq -\frac{1}{2}\overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}} \overline{\boldsymbol{x}}_j + \overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x}_0.$$

Since the left hand side is equal to the right hand side (up to the population index), we could study the discriminants:

$$-\frac{1}{2}\overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}} \overline{\boldsymbol{x}}_j + \overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}} \boldsymbol{x}_0$$

consisting of

- a constant coefficient (scalar): $-\frac{1}{2}\overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}} \overline{\boldsymbol{x}}_j$;
- a linear coefficient vector: $\overline{\boldsymbol{x}}_j^T \boldsymbol{S}^{-1}_{\text{pooled}}$.

Table 2: Discriminant analysis:Percentage of well classified observation in each drug class.

	Anti-	Anti-				
	depressants	psychotics	Anxiolitics	Hypnotics	Placebo	Stimulants
Antidepresants	58.06	3.23	0.00	0.00	25.81	12.90
Antipsychotics	0.00	43.75	0.00	0.00	56.25	0.00
Anxiolitics	0.00	0.00	12.50	0.00	62.50	25.00
Hypnotics	0.00	5.00	5.00	30.00	60.00	0.00
Placebo	5.33	1.33	0.00	4.00	88.67	0.67
Stimulants	3.13	3.13	0.00	0.00	40.63	53.13

In exactly this way, the SAS output for the linear version is presented.

```
                               Linear Classification Rule

                                Linear Discriminant Function

                             _     -1 _                                  -1 _
              Constant = -.5 X' COV   X  + ln PRIOR     Coefficient = COV   X
                              j        j           j    Vector               j


                           Linear Discriminant Function for class

Variable    Label       antidep     antipsy      anxiol      hypnot     placebo      stimul

Constant                  -1389       -1376       -1421       -1346       -1359       -1404
TAW_min1    AW min      4.47684     4.50787     4.62369     4.34552     4.43459     4.54279
TQW_min1    QW min      4.08040     4.15691     4.19095     3.97479     4.06276     4.11706
TSWS1_min1  SWS1 min    4.01093     4.06067     4.10678     3.88604     3.97078     4.03787
TSWS2_min1  SWS2 min    4.23988     4.24359     4.34992     4.10170     4.17790     4.25670
TIS_min1    IS min      4.25856     4.83064     4.56278     4.77918     4.32945     4.08450
TPS_min1    PS min      3.54727     3.62606     3.76478     3.50182     3.65985     3.69319
TAW_min2    AW min      0.87439     0.76345     0.74964     0.95291     0.85678     0.84124
TQW_min2    QW min      0.75161     0.54048     0.61318     0.77328     0.67167     0.67715
TSWS1_min2  SWS1 min    0.87459     0.70087     0.77185     0.94570     0.84544     0.86229
TSWS2_min2  SWS2 min    0.56249     0.51111     0.41581     0.63517     0.59102     0.55795
TIS_min2    IS min      2.03727     2.42074     2.12161     2.18998     2.23337     2.08833
TPS_min2    PS min      0.64477     0.66075     0.48419     0.76533     0.55725     0.62062
```

Thus, for a new observation, the six discriminants are evaluated, and the observation is assigned to the group with the largest value.

Table 2 shows the percentage of well classified observation for the linear version.

Table 3: Discriminant analysis:Percentage of well classified observation in each drug class.

	Anti- depressants	Anti- psychotics	Anxiolitics	Hypnotics	Placebo	Stimulants
Antidepresants	87.10	3.23	0.00	0.00	3.23	6.45
Antipsychotics	0.00	100.00	0.00	0.00	0.00	0.00
Anxiolitics	0.00	0.00	100.00	0.00	0.00	0.00
Hypnotics	0.00	0.00	0.00	100.00	0.00	0.00
Placebo	6.67	2.00	0.00	12.67	67.33	11.33
Stimulants	0.00	0.00	0.00	0.00	9.38	90.63

```
                                 The DISCRIM Procedure
                        Classification Summary for Calibration Data: WORK.TRAINT
                       Resubstitution Summary using Linear Discriminant Function

                                Error Count Estimates for class

                antidep     antipsy      anxiol      hypnot     placebo      stimul       Total

Rate             0.3548      0.4375      0.7500      0.6000      0.0867      0.3750      0.4341
Priors           0.1667      0.1667      0.1667      0.1667      0.1667      0.1667
```

We observe that all only 58.06 % of the antidep are correctly classified; that the remains are classified into other classes (3.23 % into antipsy, 25.18 % into placebo and 12.0 % into Stimul). For the rest of the drug classes, in can be seen also how are they classified.

The overall (total) error rate is computed as follows:

$$\begin{aligned} 0.3548 \times 0.1667 + 0.4375 \times 0.1667 + 0.7500 \times 0.1667 & \\ +0.6000 \times 0.1667 + 0.0867 \times 0.1667 + 0.3750 \times 0.1667 &= 0.4341, \end{aligned}$$

a very poor result. It can be explained by the fact that we are using a summary measured of the data that it was also shown in previous section that it is not always a good choice. As this is used only for illustration purpose, it is also important to remark that the data we start with is also playing an important role for the classification of the drugs.

Quadratic Version Since the quadratic discriminant function is not as easy to decompose as the linear discriminant function, the first part of this output panel is not included in SAS. One only gets a summary of the performance of the classification (Table 3).

```
                                 The DISCRIM Procedure
                    Classification Summary for Calibration Data: WORK.TRAINT
                   Resubstitution Summary using Linear Discriminant Function

                             Error Count Estimates for class

                 antidep    antipsy     anxiol     hypnot    placebo     stimul      Total

Rate              0.1290     0.0000     0.0000     0.0000     0.3267     0.0938     0.0916
Priors            0.1667     0.1667     0.1667     0.1667     0.1667     0.1667
```

The rule is different, the classification performance is also different, if we compare it with the previous result, the overall error rate is much lower than the previous analysis. From the previous results we can see that a quadratic discriminant function may be a better choice.

8.2 Classification Tree Analysis

The data used in the previous section will be analyzed using a non-parametric technique (classification tree analysis). In the construction of the classification tree several screening status can be used, for this particular example the so called Gini diversity index was used (Therneau and Atkinson 1997). This index measures the "impurity" (i.e., heterogeneity) in the node and it belongs to the interval $[0, 1]$. Breiman *et al.* (1984) recommended that binary splits be chosen to minimize the Gini diversity index. In this particular example we do not have missing observation, thus no surrogates split were necessary. In order to find the best split, a 10-fold cross-validation was used, the tree with the minimal cross validation relative error was selected. Figure 3 shows the cross validation relative error, reaching the minimal value for a tree of size 14, which correspond to a cost complexity of 0. The final tree is shown in Figure 4. The top node contain the entire sample, each of the remaining nodes contains a subset of the sample in the node directly above it. Furthermore, any node contains the sum of the samples in the nodes connected to and directly below it. The tree thus splits samples. Each node can be thought of as a cluster of objects (cases) which is to be split by further branches in the tree. The numbers below the terminal nodes show how the cases are classified by the tree into a particular drug class. Tree prediction models add two ingredients: the predictor and predicted variables labeling the nodes and branches. It can be seen that the first split is given by TPS_min1, which in the discriminant analysis is the most important variable. $TPS\_min1 < 45.9$ splits the root node. Cases meeting this criterion move left to node 2; the rest move right to node 3. The procedure continues further and we end up in the final nodes. Table 4 shows the percentage of well classified observation when the classification tree is used.

It can be seen that the performance of this method is similar to the linear discriminant analysis, and it can be also confirmed by the classification error obtained.

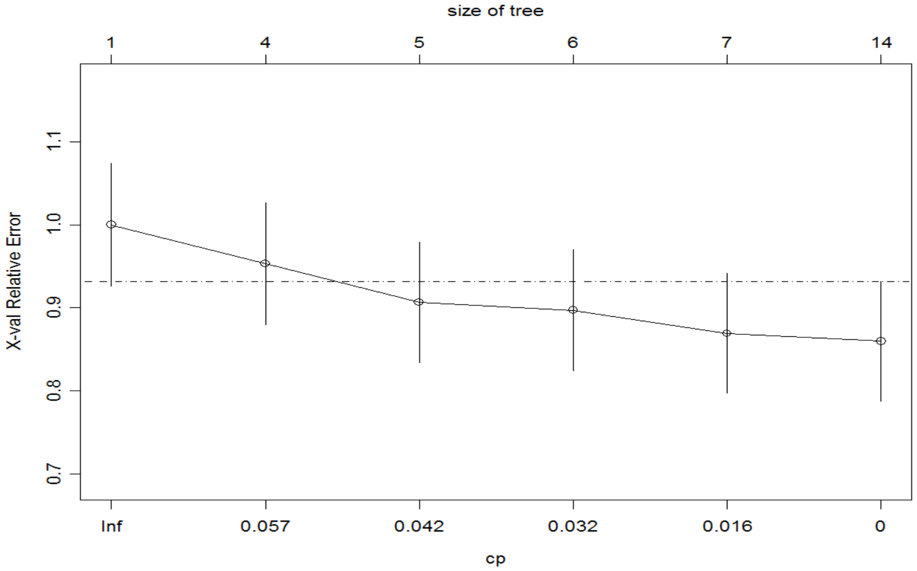

Figure 3: The X-val relative error in function of cost complexity parameter and the size of the tree.

```
                                 The classification tree Procedure

                             Error Count Estimates for class

                 antidep     antipsy     anxiol      hypnot     placebo      stimul       Total

Rate              0.3871      0.5625     1.0000      0.1500      0.1267      0.3750      0.4336
Priors            0.1667      0.1667     0.1667      0.1667      0.1667      0.1667
```

It can be seen that the linear discriminant analysis and the classification tree produce similar results in terms of well classified observation as well as overall error rate.

8.3 Some Reflections on Discrimination and Classification Methods

Several classification techniques can be applied in order to classify a compound. It can be used parametrically or in a non-parametric formulation of the problem. In case that a parametric technique (in this case we discussed discriminant analysis) is used, it is important to explore if the discriminant function to be used is linear or quadratic. For this particular example, the quadratic version seems to perform better than the linear version. Another important aspect to take into account is the data we are using to classify the compounds. If the data we start with is not giving us a good idea about the behaviour of the compound, in term of the changes in sleeping stages, it may be a good idea to first model the behavior, using linear or nonlinear mixed effects model in order to get a better summary of the data, and use it as an input in the classification procedure. For example,

Table 4: Classification tree analysis:Percentage of well classified observation in each sleeping stage.

	Anti- depressants	Anti- psychotics	Anxiolitics	Hypnotics	Placebo	Stimulants
Antidepresants	61.29	3.23	0.00	6.45	29.03	0.00
Antipsychotics	12.50	43.75	0.00	37.50	6.25	0.00
Anxiolitics	0.00	0.00	0.00	37.50	62.50	0.00
Hypnotics	0.00	0.00	0.00	85.00	15.00	0.00
Placebo	1.33	0.67	0.00	6.00	87.33	3.33
Stimulants	15.63	3.13	0.00	9.38	9.38	62.50

if we assumed that within a drug class, the changes in sleeping stages should be similar, we can try to model the class using the profiles of the rats, and then use the parameters in the model as an input datafile for the classification procedure.

There are a large number of methods that an analyst can choose from when analyzing classification problems. Tree classification techniques, when they “work” and produce accurate predictions or predicted classifications based on few logical if-then conditions, have a number of advantages over many alternative techniques. Simplicity of results. In most cases, the interpretation of results summarized in a tree is very simple. This simplicity is useful not only for purposes of rapid classification of new observations (it is much easier to evaluate just one or two logical conditions, than to compute classification scores for each possible group, or predicted values, based on all predictors and using possibly some complex nonlinear model equations), but can also often yield a much simpler “model” for explaining why observations are classified or predicted in a particular manner.

Tree methods are nonparametric and nonlinear. The final results of using tree methods for classification can be summarized in a series of (usually few) logical if-then conditions (tree nodes). Therefore, there is no implicit assumption that the underlying relationships between the predictor variables and the dependent variable are linear, follow some specific non-linear link function, or that they are even monotonic in nature. For example, some continuous outcome variable (Y_1) of interest could be positively related to another variable (X_1) if X_1 is less than some certain amount, but negatively related if it is more than that amount (i.e., the tree could reveal multiple splits based on the same variable X_1, revealing such a non-monotonic relationship between the variables). Thus, tree methods are particularly well suited for data mining tasks, where there is often little a priori knowledge nor any coherent set of theories or predictions regarding which variables are related and how. In those types of data analyses, tree methods can often reveal simple relationships between just a few variables that could have easily gone unnoticed using other analytic techniques.

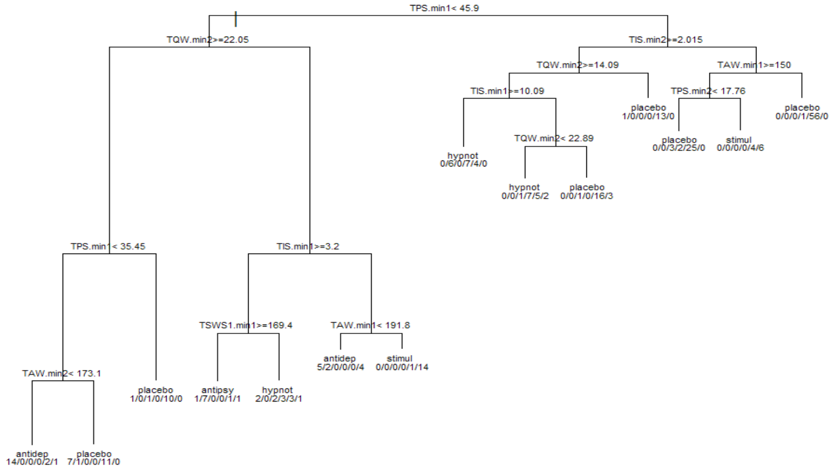

Figure 4: The final tree with the number of observation classify in a particular drug class.

9 References and Bibliography

Abrams, K., Jones, D.R., Sheldon, T.A., Song, F., Sutton, A.J. (2000). *Methods for Meta-analysis in Medical Research*. New York: John Wiley.

Aerts, M., Geys, H., Molenberghs, G., and Ryan, L.M. (2002). *Topics in Modelling of Clustered Binary Data*. London: Chapman & Hall.

Agresti, A. (1990). *Categorical Data Analysis*. New York: John Wiley.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and regression trees*. New York, Chapman and Hall.

Breslow, N.E. and Day, N.E. (1990). *Statistical Methods in Cancer Research (Vol. 1): The Analysis of Case-Control Studies*. Lyon: World Health Organization.

Celeux, G. and Nakache, J.P. (1994). *Analyse discriminante sur variables qualitative.* Paris, Polytechinica.

Cohen, D. and Cailloux-Cohen, S. (1995). *Guide critique des médicaments de lâme*. Québec, Les Editions de lHomme.

Costentin, J. (1993). *Les médicaments du cerveau*. Paris, Odile Jacob.

Cox, D.R. and Hinkley, D.V. (1990). *Theoretical Statistics*. London: Chapman & Hall [Background on the inferential and theoretical aspects of the approaches discussed].

Deniker, P. (1982). Vers une classification automatique des psychotropes à travers un fichier informatisè de leurs propriétés. *Annales Médico-psychologiques*, **1**, 25–27.

Diggle, P.J., Heagerty, P., Liang, K-Y., and Zeger, S.L. (2002). *Analysis of Longitudinal Data*. New York: Oxford University Press.

Dunn, G. and Everitt, B. (1995). *Clinical Biostatistics*. London: Arnold.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Heidelberg: Springer-Verlag.

Fisher L., van Ness J.W. (1973). Admissible Discriminant Analysis. *Journal of American Statistical Association*, **68**, 603–607.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. **7**, 179-188.

Johnson, R.A. and Wichern, D.W. (1992). *Applied Multivariate Statistical Analysis.* London: Prentice-Hall [Presents a multitude of concepts and methods of multivariate analysis].

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data.* New York: John Wiley and Sons.

Kleinbaum D.G. (1996). *Survival Analysis, A Self-Learning Text*. New York: Springer-Verlag.

Krzanowski, W.J. (1988). *Principles of Multivariate Analysis - A User's Perspective*. Oxford Science Publications.

Laird, N.M. and Ware, J.H. (1982). Random effects models for longitudinal data. *Biometrics* **38**, 963–974.

Le, C.T. and Boen, J.R. (1995). *Health and Numbers*. New York: Wiley-Liss.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

McLachlan, G.J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. New York, John Wiley and Sons.

McLachlan, G.J. and Peel, D. (2000). *Finite mixture models*. New York: John Wiley.

Neter, J., Kutner, M.H., Nachtsheim, C.J. and Wasserman, W. (1996). *Applied Linear Statistical Models, 4th Ed.* New York: McGraw-Hill.

Pagano, M. and Gauvreau, K. (1992) *Principles of Biostatistics*. Belmont, CA: Duxbury Press.

Pinheiro, J.C. and Bates, D.M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4**, 12–35.

Oughourlian, J. M. (1984). *La personne du toxicomane. Psychosociologie des toxicomanies actuelles dans la jeunesse occidentale*. Toulouse, Privat.

Rothman K.J. and Greenberg, S. (1998). *Modern Epidemiology*. Philadelphia: Lippincott-Raven Publishers.

Rosner, B. (1994). *Fundamentals in Biostatistics* (4th ed). Belmont, CA: Duxbury Press.

Saporta, G. (1990), *Probabilités Analyse des Données et Statistique*, Paris, Editions Technip.

Shoukri, M.M. and Pause, C.A. (1999). *Statistical Methods for Health Sciences* (2nd ed). Boca Raton: CRC Press.

Therneau, T. and Atkinson, E. (1997). An introduction to recursive partitioning using the rpart routines. *Technical report*.

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*, Springer Series in Statistics, Springer-Verlag, New-York [Overview of linear mixed models for longitudinal data, with a lot of examples].

Zarifian, E. (1988). *Les jardiniers de la folie*. Paris, Odile Jacob.

Zarifian, E. (1996). *Le prix du bien-être. Psychotropes et société*. Paris, Editions Odile Jacob.